Politecnico di Milano

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING
Master of Science – Management Engineering

POLITECNICO
MILANO 1863

Automatic detection of Overfunding in reward-based crowdfunding campaigns: a classification-based approach

Supervisor
Prof. Vincenzo Butticè

Co-Supervisors
Prof. Cristina Rossi-Lamastra
Prof. Mara Tanelli

Candidates
Francesca Credentini – 944392
Aurora Finotto – 946350

Academic Year 2020 – 2021

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

Abstract

Historically, crowdfunding and the dynamics of success have been extensively studied in literature. However, recently, one of the crowdsourcing unexplored areas, namely the overfunding, has started to arouse interest in scholars' papers. With our research, we filled some *"literature gaps"*, offering important insights for future studies and with relevant implications both from a theoretical and managerial perspective. We used a machine learning approach, specifically relying on the *Classification Learner app* provided by MATLAB to conduct several analyses, offering an innovative methodological approach in investigating the dynamics of overfunding. Our research proposes to offer a clear and comprehensive definition of the phenomenon, trying to understand which, among static and dynamic variables, are the most relevant in achieving and predicting the *"over-success"*. The study aims also to understand if, as in the success case, a critical time horizon exists which is crucial to predict the phenomenon. The analysis sample included 157 campaigns able to reach the final target before their deadline, all launched and concluded between 2016-2017 on Kickstarter. To develop the models, we considered 10 static predictors and 4 dynamic predictors. With our findings we managed to confirm that, for a high-overfunding threshold, it is possible to discriminate between success and overfunding. Concerning the relative importance of static predictors, campaign-related and fundraisers-related ones are crucial, while among the dynamic factors, the percentual pledge over funding target is the most relevant in impacting the phenomenon. Regarding the time horizon, to reach the best accuracy possible of the model, and to have a comprehensive view of the overfunding, 31 days should be considered. Our work could be generalised by future researches, collecting larger samples and considering different crowdfunding contexts or platforms.

Estratto

Storicamente, il crowdfunding e le dinamiche del successo sono stati ampiamente studiati in letteratura. Tuttavia, di recente, una delle aree inesplorate del crowdfunding, ossia *l'overfunding*, ha iniziato a suscitare l'interesse degli studiosi. Con la nostra ricerca abbiamo colmato alcune *"lacune letterarie"*, offrendo importanti spunti di approfondimento per studi futuri, con rilevanti implicazioni sia dal punto di vista teorico che manageriale. Abbiamo utilizzato un approccio di machine learning, adoperando, in particolare, la *Classification Learner app* fornita da MATLAB, per condurre diverse analisi, fornendo così un approccio metodologico innovativo per investigare le dinamiche di *overfunding*. La nostra ricerca si propone di trovare una definizione chiara e completa del concetto di *overfunding*, analizzando quali tra le variabili statiche e dinamiche siano le più rilevanti per raggiungere e prevedere *"l'over-successo"*. Lo studio mira anche a capire se, come nel caso del successo, esiste un orizzonte temporale cruciale che deve essere considerato per prevedere il fenomeno. Il campione di analisi comprende 157 campagne che hanno raggiunto il target prima della scadenza, tutte lanciate e concluse tra il 2016 e il 2017 su Kickstarter. Per sviluppare i modelli, abbiamo considerato 10 variabili statiche e 4 dinamiche. Con i nostri risultati siamo riuscite a confermare che fissando una soglia elevata di *overfunding*, è possibile discriminare tra successo e *over-successo*. Per quanto riguarda l'importanza relativa dei predittori statici, i fattori relativi alla campagna e al *fundraiser* risultano essere cruciali, mentre tra quelli dinamici, la percentuale di investimenti raggiunti sul target richiesto è il più rilevante per predire il fenomeno. In ultima istanza, relativamente all'orizzonte temporale, per poter raggiungere la massima accuratezza possibile del modello ed avere una visione globale del fenomeno, si dovrebbero considerare 31 giorni. Il nostro lavoro potrebbe essere generalizzato da ricerche future, adoperando campioni più grandi e considerando diversi contesti o piattaforme di crowdfunding.

Executive Summary

State of the art and research objectives

Today, crowdfunding is a widespread and well-established phenomenon, always arousing the interest of many scholars, who have proposed to study its most important dynamics. In particular, the majority of the literature on crowdfunding has been widely committed to investigating the determinants of success of a crowdfunding campaign (Butticè et al., 2018).

Over the years, many features influencing the success of a campaign have been investigated. The literature has shown that both static factors, namely those present at the launch of the campaign, and dynamic factors, which can be modified during the life-cycle of the project, are relevant to achieve a successful outcome. Moreover, Colombo and colleagues (2015) resumed different analysis, and revealed the importance of the performance of the first week for the campaign to be successful, precisely finding that the percentage of the target funded after one week can be considered as an indicator of the probability of success.

Despite the huge interest existing among scholars for both crowdfunding in general and the success dynamics in particular, there is a peculiar and recent phenomenon which has been little studied in literature, namely overfunding.

Specifically, one of the topics treated is overfunding as a source of market inefficiency, since it can bring to a suboptimal allocation of resources towards highly visible campaigns (Li et al., 2020), thus becoming the main failure cause of underexposed projects. This dynamic has been considered as a deterrent for aspiring entrepreneurs from fundraising because of the fear of being overshadowed, which in turn might reduce the crowdfunding platform into a *"market for lemon",* excluding high-quality projects (Li et. al., 2020, Koch et al., 2018).

Also, negative implications for fundraisers have been analysed. The majority of scholars agree that extra funds may create inefficiencies (Makýšová L. et al., 2017; Li Y. et al., 2020; Liu F. et al., Koch J. et al., 2020; Svatopluk Kapounek, Zuzana Kučerová, 2019) since project creators may not be able to manage the excess of resources thus failing in delivering the rewards on time and, from a psychological perspective, this could inflate their egos, spurring them to take on unnecessary risks (Li et al., 2020). Some studies also focus on the main determinants affecting the overfunding, analysing some campaign-related factors such as the funding target (Ma X et al., 2018) and some fundraisers-related factors like gender (Cicchiello A. F. et al., 2020).

From the literature review many unexplored areas emerged, since the study of the phenomenon is still in its early stages, with few, very recent, contributions. Firstly, the literature lacks an integrated definition and operationalization of the overfunding concept: the papers selected often treat alike a campaign that has received a few dollars more than the target and another that collected ten times the initial goal, without clearly discriminating among success and overfunding, making the boundaries between one and the other very blurred. We argue that these considerations may prove to be over simplistic. For this reason, our first research objective focuses on the possibility to provide a comprehensive and concrete definition of overfunding, taking a step forward from existing literature, which remains very vague on the subject.

In addition, our analysis aims to investigate whether, as in the case of success, it is possible to predict overfunding based on certain static and dynamic characteristics, and which among the above mentioned are the most relevant to predict the phenomenon.

As a last point, the time horizon is taken into account. For success, it was discovered that the first week is crucial to determine the outcome of the project and, on the same line, we want to investigate which moment during the duration of the campaign is the most critical for the overfunding phenomenon.

Methodology

To investigate our research objectives, we resorted to machine learning techniques. Several classification models exist: *perceptron-based techniques, statistical learning algorithms, support vector machines,* and *logic-based algorithms*. Classification models are trained on a training set and subsequently validated on a test set.

In order to perform the analysis, we built a database including variables of different nature. Precisely, we started using a sample including 352 campaigns launched on Kickstarter between 2016 and 2017, consisting of 195 unsuccessful campaigns and 157 projects that, instead, have reached the target amount within the end of the campaign. Then, we focused only on the latter (157 instances), in order to build an empirical model able to discriminate between success and overfunding, and to do so, we employed a set of classifiers derived from machine learning techniques. Our classifiers initially considered 20 static and 4 dynamic predictors, chosen on the basis of the factors that crowdfunding literature has shown as influencing the success of a campaign. But, given the limited size of our dataset and the high risk of incurring in overfitting, we reduced, through a correlation analysis, the features to 14, namely 10 statics and 4 dynamics, avoiding to consider the redundant ones.

In the empirical model, the response variable chosen was the predicted class label associated to a campaign, which is determined by the classification models according to the predictors given as input. In our case, the response variable was the *overfunding class*, namely whether the campaign is *"simply"* successful or *over-successful*. In order to clearly distinguish among the twos, we considered a threshold to define overfunding that, if overcome, it means that the campaign is *over-successful*, otherwise it is just successful. The threshold value is calculated as *pledge/target* and varied in a range between 120% and 175% with a deviation of 5%. This procedure has been done with the willingness to assess which among these percentages is the most representative of the overfunding phenomenon in our empirical model. At this point, we created a total of 32 databases, 1 static and 31 dynamics, for each of these percentages. Precisely, we decided to consider a time span of 31 days (closing date of the majority of the campaigns) to have a complete spectrum of the overfunding phenomenon.

In order to perform the machine learning classification, we divided our sample in the training set and in the test set. The former includes 70% of the total instances (110 campaigns), the latter includes the remaining 30% (47 campaigns). The last 47 campaigns have been used during the testing phase, in order to assess the behaviour of the classification models previously trained on the training set, when new instances are provided. For this process, we employed the MATLAB *Classification Learner app*, training 15 possible classifiers. Then, we selected the three most accurate classifiers: *Coarse Tree, Ensemble RUSBoosted Trees* and *Ensemble Bagged Trees.* We obtained a total of 1152 classification trained models: 32 models (1 for the static dataset and 31 for the dynamic ones) for each of the three best classification algorithms resulted from the training phase (*Coarse Trees, RUSBoosted Trees, Bagged Trees)*, for each overfunding percentage selected (120%, 125%, 130%, 135%, 140%, 145%, 150%, 155%, 160%, 165%, 170%, 175%). We tested their performances using the test set and we computed four indicators: *accuracy, precision, recall* and *F1 score*.

The analysis proceeded with the computation of the importance of the predictors for each of the three best classification models. To this end, we adopted both *Filter* and *Wrapped* methods. The latter consists in the computation of predictors' importance with the function *PredictorImportance* provided by MATLAB. To further validate our findings, we also evaluated how the accuracy of the model changed by excluding one static predictor at a time.

Results and conclusions

The results of our empirical model led us to several conclusions.

As developed in the *Methodology*, we fixed an overfunding percentage threshold, calculated as the ratio between the pledge and the target amount, among 120% and

175%, trying to understand which of these percentages allowed our classification tool to identify the phenomenon of overfunding with the highest accuracy possible. We found that 175% was the most accurate for our classifiers, demonstrating that for a high-overfunding threshold, success and overfunding are two clearly distinguished conditions. Specifically, our empirical model provides a possible operationalisation that could be used to define and differentiate the two phenomena, which directly emerge from the confidence level through which the classifier is able to discriminate among success and overfunding. Finally, we conclude this section embracing the definition provided by Mollick (2014): *"Overfunding is especially used when a project's funding is considerably higher than its funding goal",* also managing to concretely define the general concept of *"considerably higher"* into mathematical terms. Indeed, considering our results, we believe that research on overfunding should not only consider *if* a campaign has reached its target but also *how much* money it has been collected beyond the target itself.

The results of the analysis of the predictors' relative importance revealed that when only static predictors are considered, the most important ones are campaigns-related predictors, such as target, and project category and fundraisers-related factors, such as the fundraiser being a serial one, and whether the project creator is a single individual or a team. When we included in the model dynamic predictors, the most important one is the percentage of funding target pledged until that moment. These findings are consistent in all the three methods employed (*Filter, Wrapped, Excluded Predictors*).

Furthermore, with regard to the temporal impact on the probability of overfunding, our work notes that for this phenomenon it is important to have a wider time spectrum than success, in which it was sufficient to consider only the first week. Indeed, to have an overall view of overfunding and achieve high accuracy of the model (98,2%), all the 31 days of the campaign should be taken into account. At the same time, paralleling the success, our work corroborates the informative importance of the first week to predict the outcome of the campaign. Indeed, also overfunding can be predicted with a fair degree of approximation considering only that week.

Our work has several theoretical and managerial implications. For the former, we provided a comprehensive definition and operationalization of the overfunding concept, so to discriminate in a clear way between success and "*over-success*", deepening a concept in which literature still remains vague. In addition, our study provided a contribution by analysing the relative importance of 14 static and dynamic variables, and by making a comparison with the factors influencing success. These broader spectrums of factors considered, provide a more integrated view of the impact that the choices undergone by fundraisers have on the *"over-success"* of their campaigns. Also, we derived important insights in the time span that has to be considered in the overfunding case, while developing a parallelism between success and overfunding, revealing that the first week could be used in predictive models that aim to compare the two phenomena, with a more than good confidence level. Furthermore, we provide a methodological contribution. In fact, to carry out our analysis, we applied machine learning classification models to the overfunding context, contrary to the statistical and qualitative techniques used until this year in the literature. Finally, our work has important practical implications both for project creators, providing them with the most important predictors of overfunding, and for crowdfunding platforms, who could adopt our innovative tool and create a dedicated section on the platforms to help campaigns' fundraisers to assess the expected outcome of their projects from the launch of the campaign until its end.

While being very insightful on the overfunding phenomenon, our study has some limitations and gaps which could be covered by future researches. First of all, for our analysis we used a small sample (157 campaigns), so future researches have to corroborate and validate our findings using a wider dataset. Additionally, the current study is limited to the analysis of campaigns launched and concluded between 2016 and 2017, so our results have to be generalised considering more recent campaigns.

Moreover, we evaluated a very specific crowdfunding setting (reward–based context, Kickstarter platform), so future studies should investigate different crowdfunding models.

Chapter 1 – Introduction

Despite the interest in crowdfunding in recent years has led many researchers to investigate the crowdfunding context, focusing in particular on success dynamics, there is still a paucity of study regarding the overfunding phenomenon. These studies touch only few topics, such as the fact that overfunding could lead to market inefficiencies, reducing the crowdfunding environment to a *"market for lemon"*, and on some determinants affecting the phenomenon (e.g., target, gender).

Our work contributes to the existing literature addressing three main objectives. We aim at advancing the knowledge on this theme, through the development of an empirical model relying on machine learning techniques. For all the analysis carried out we employed the *Classification Learner app* provided by MATLAB.

Firstly, the Dissertation proposes to find a comprehensive definition and operationalisation of overfunding concept: we identified 12 overfunding percentage thresholds calculated as *pledge/target* (ranging from 120% to 175%), and we trained the classification model using all the different datasets created, 1 static and 31 dynamics, for each of the percentage defined. Conducting this analysis, we made a step forward from existing literature, and considering the best threshold in terms of accuracy performances for our classifiers (as in the case, 175%), we stated that for a high-overfunding threshold, success and overfunding are two clearly distinguished conditions.

Secondly, we aim to assess the relative importance of different static and dynamic factors on the overfunding phenomenon, providing several insights. We relied to *Filter* and *Wrapped* Methods to obtain a ranking of static predicting factors: the importance scores highlighted that, campaign-related features (funding target and project categories), and, fundraisers related factors (such as whether the fundraiser is a serial one, and whether the project creator is a single individual or a team) are the most important in predicting an overfunding outcome. Moreover, among the dynamic factors, we found that the percentual pledge over funding target is the most important one (also in comparison to static factors).

Lastly, the time horizon has been taken into consideration, since extant studies never analysed which moments whiten the campaign duration is the most critical to predict and achieve the overfunding phenomenon. We found that for overfunding the best time horizon to be considered, both in terms of completeness and accuracy of the model, is 31 days. However, a good level of approximation to predict the phenomenon may be obtained also considering just the first campaign week, consistently with what happens in the success domain.

We thoroughly believe that our thesis has opened the door to new studies and researches, in order to investigate other interesting areas about overfunding, deepening the theme with new theoretical results, using machine learning algorithms. Moreover, our findings have important practical implications both for project creators and for crowdfunding platforms, that could improve the efficiency of the platform itself by adopting our model.

The rest of the work proceeds as follows. In the second chapter, we offer a comprehensive overview of the extant literature about the general context of crowdfunding, success dynamics and overfunding. In the third chapter, we develop our three research objectives. In the fourth chapter, we provide a theoretical framework of machine learning classification models. In chapter five and six, we explain the methodology. In the seventh chapter, we discuss our results, highlighting the implications of our study, and assessing its limitations, which open avenues for future researches.

Chapter 2 – Literature review

The following chapter aims to define the determinants and the dynamics of the overfunding previously studied by scholars. To this end, we performed a review of the extant literature about the topic.

The chapter proceeds as follows: the first section discusses the methodology employed for searching the papers used to assess the extant literature on overfunding. Then, considering that the latter is a very recent and not yet explored phenomenon, our analysis has started with general considerations about the crowdfunding context and its main elements. In the third section, we present all the factors that the literature identifies as determinants of the success of a crowdfunding campaign, in particular we focus on campaign, fundraisers and backers-related factors. Moreover, we dedicated a specific section firstly to linguistic features, then to dynamic variables, which have both been investigated as factors influencing crowdfunding result as well. In the fourth and final section, we address the hearth of our dissertation theme, namely overfunding, specifying the main areas of interest presented by scholars.

2.1. Methodology

We performed the literature research between the beginning of December 2020 and the middle of March 2021. The objective was to collect and summarize the existent literature about the determinants and the dynamics of the overfunding and to provide a comprehensive framework about this phenomenon. This paragraph explains the methodology applied for conducting the literature research and analysis about the overfunding topic.

2.1.1. The overfunding phenomenon

To review the extant knowledge about the dynamics of the overfunding and its main determinants, we have conducted systematic researches, and we firstly employed the

Scopus database[1], which combines a comprehensive, expertly curated abstract and citation database with linked knowledge and data produced by scholars across a wide variety of fields. Indeed, as reported by Emily Glenn, Associate Dean, *"Scopus really plays a vital role in helping our researchers, particularly early investigators, better understand the scholarly communications landscape."* In particular, using this database we performed the following steps: keywords identification and documents' selection.

Keywords identification

The first stage of the research consisted in the definition of the keywords. Since the purpose of our study was to provide a comprehensive investigation about the crowdfunding dynamics that cause overfunding, we decided to divide our query in two parts: a first one related to the identification of the crowdfunding context in general and a second one specifically related to the overfunding.

On the Scopus database, for the first part, the terms *"crowdfunding", "fundraising", "crowdsourcing", "crowd-funding", "fund-raising"* and/or *"crowd-sourcing"* have been selected; with respect to the second part**,** the term *"overfunding"* has been adopted, in order to obtain a complete framework of the main documents investigating the overfunding. The selected keywords have been searched within *Title*, *Abstract* and *Keywords*. As overfunding is a recent topic of interest, we decided not to include any constraints on the language and year of publication, so to not reduce the scope of the analysis and maintain a broader perspective.

The query string for the Scopus research resulting from these considerations was the following:

TITLE-ABS-KEY (("crowdfunding" or "fundraising" or "crowdsourcing" or "crowd-funding" or "fund-raising" or "crowd-sourcing") AND ("overfunding"))

[1] https://www.scopus.com/home.uri

Document's selections

The aforementioned query produced 23 documents on Scopus, on which we have carried out the following operations:

1) We exported an excel sheet with all the documents and we created a table containing the following information: *Title, Author(s), Year of publication, Source title, Abstracts, Authors Keywords, Publisher, Document type.*
2) Then we proceeded to the phase of *abstract reading,* to discard those not in line with the scope of our research. Thanks to this analysis, we managed to discard 7 papers, that were not focused precisely on the overfunding phenomenon.
3) Following, we started reading the full text of the selected documents (*text reading phase*) and at the end we considered 12 documents useful for our work, as they focus on proposing an overview of overfunding phenomenon and to analyse its dynamics.
4) Lastly, we categorized the various papers by relevance to identify those from solid and famous sources, employing the *Scimago Journal Rank* or *SJR indicator*[2], that measures the degree of scientific influence of academic journals. In this way we have identified the quartiles of relevance for all the papers, divided in: *Q1: Best Case; Q2: acceptable case; Q3* and *Q4 of minor/non-relevance*.

Overview of the selected documents

Hereby, we provide a brief overview of the relevant documents from the search on Scopus. We have found the following document types:

- Article (6)
- Conference Papers (6)

All the documents have been published between 2021 and 2016: most of them are dated back to 2020 (46%) and 2018 (31%), which suggests the recent interest of scholars in the overfunding phenomenon and, subsequently, the need to systematize

[2] https://www.scimagojr.com/

the extant knowledge. In particular, the papers we analyzed, focus on the negative implications of overfunding for all crowdfunding stakeholders and on the dynamics leading to this phenomenon.

With respect to the type of crowdfunding on which the documents focus, the majority (6 documents) examines the reward-based models, and the remaining 6 investigates the overfunding phenomenon exclusively for one crowdfunding model, such as exclusively for the equity crowdfunding (4), donation (2). Moreover, in our analysis sample, there are 4 papers that analyze specific national markets, but we decided not to exclude them as they are valid to provide a complete overview on the phenomenon. Indeed, since the articles and the conference proceedings resulting from our Scopus research were very recent and limited in number, we preferred to keep them all, in order to have a complete and clear vision of the literature state of the art.

This reasoning was applied even after the *SJR index* was checked. In particular, only two articles revealed to be considered *"very good"* in the *Q1 range*, the others a lower step. Despite the decision to consider them in the analysis, we have taken a careful look, trying to filter out only useful and reliable information from these papers.

Forward citations analysis

After the aforementioned investigations, we performed the forward citations analysis based on the 12 papers previously selected. To do so, we adopted the function *"cited by"* [3]provided by Scopus and, for each article, we exported in a table the following information about the papers: *Authors, Year, Title, Source title, Abstract* and *Keywords* and *Document type.* Moreover, we added the column *"Citation ID"*, identifying which of the Scopus articles previously read were cited by the paper of the forward table. After we completed this collection procedure for each of the 12 articles, we decided

[3] https://blog.scopus.com/posts/cited-reference-searching-in-scopus

not to consider the new papers found, since they were not in line with the scope of our analysis.

Document's selection from additional sources

As a last step, we implemented a search on *Google Scholar* [4] to have a broader overview of the overfunding phenomenon.

The procedure adopted to define the query string was the same used for Scopus: a first part related to the identification of the crowdfunding context in general and a second one specifically connected to the overfunding. For the Google Scholar research, with respect to the first part, we used *"crowdfunding"*, while for the second one, the word "*overfunding"* was employed.

In order to obtain as a result of the search only the documents that contained the keywords in the title tag, the command *"Allintitle"* [5]was adopted, so to find papers focused on the theme. Here, the string used in Google Scholar:

Allintitle: (crowdfunding | overfunding)

Among the documents obtained with this methodology, we removed the ones already found through Scopus and Scopus forward. At the end of this process, 10 papers were kept and then, through the *abstract reading phase*, we excluded 3 of them. On the other hand, 7 articles succeeded the *full text reading phase* and resulted relevant for the scope of our search. After that, as in Scopus, we performed the forward analysis for these 7 articles.

We found the following document types:

- Article (2)
- Thesis' chapters (2)

[4] https://scholar.google.it/
[5] https://moz.com/learn/seo/search-operators

- Conference Papers (3)

As in Scopus, all the documents are recent, dated between 2015 and 2019, which allowed us to make the same considerations as in the previous phase. Moreover, alike the analysis conducted on the other database, the majority (5 documents) examines the reward-based models, and the remaining 2 investigates the overfunding phenomenon exclusively for the equity crowdfunding model.

Finally, to complete the articles' exploration, we performed a search on *SSRN*[6] in January 2021. By entering as research query *crowdfunding and overfunding* and looking at the *Title, Abstract* and *Keywords*, the engine provided 13 results. However, among these papers, 6 had been previously found via Scopus and Google Scholar, and 7 were considered out of the boundaries of the research (e.g., related to dynamic investment policies and corporate governance).

General overview of the selected documents

Hereby, we propose an overview of the papers read, from which we extrapolated all the information to develop the literature review.

For each relevant paper analysed, for a total of 19 articles across the different citation database employed, we considered 3 factors:

- The crowdfunding model: reward, equity, lending or donation;
- The methodology applied in order to define and operationalize overfunding;
- The theories on which the different articles are based on.

A noteworthy result is that 53% of the articles takes into account the reward-based crowdfunding settings, and in these cases, the data are collected in 83% of cases from Kickstarter and 30% from both Kickstarter and Indiegogo. Also, the context of equity

[6] https://www.ssrn.com/index.cfm/en/

crowdfunding is a field which attracts a strong interest from scholars (32%), as well as the donation-based context (11%).

With respect to the methodologies employed it is interesting to note that all the articles considered adopted apply statistical and econometric models to study the overfunding phenomenon, while no document applies machine learning techniques.

2.2. Crowdfunding an overview

As anticipated at the beginning of the chapter, the main theme of our thesis is overfunding, not yet explored in depth in its most interesting dynamics. However, it is crucial to produce a preliminary introduction about the crowdfunding context and about the main elements affecting the success of a campaign, to ensure a full understanding of the phenomenon of overfunding and to clarify the terminology employed in this chapter and in the rest of the thesis.

In particular, we will explain the main crowdfunding models and we will introduce the most important actors involved, as well as their main peculiarities. Moreover, we will provide a definition of the success of a crowdfunding campaign, pointing out which are the elements influencing the latter. These considerations are derived from the review about crowdfunding and its success dynamics of Butticè and colleagues., 2018.

2.2.1 Crowdfunding: a definition

Raising initial capital is an inevitable hurdle confronting entrepreneurs at the seeding stage of their business venture (A. Schwienbacher et al., 2010). Although multiple sources of financing like banking institutions, private foundations, governmental agencies, business angels (BAs), and venture capitalists (VCs) are available for fundraising (G. Bruton et al., 2015) they are less accessible to new ventures who lack guarantees and a track of records. For this reason, crowdfunding has emerged as a financial alternative, expanding the funding opportunities for start-ups in capital markets.

To provide as a complete and general definition as possible, the one of Belleflamme and colleagues, 2014 is reported: *"Crowdfunding refers to a fundraising campaign where a creator issues an open call through the internet to the public to raise funds in the form of donations or in exchange for some reward, equity and voting rights to support initiatives for specific purposes".*

Crowdsourcing has proven to be a practical way for financing young entrepreneurs seeking early-stage fundings (Kuppuswamy and Bayus 2014). This phenomenon is an alternative fundraising-method, enabled by *Internet* and by various online platforms that connect project creators and investors, and used by entrepreneurs that exploit the union of small amounts of capital from many individuals, differently from the traditional financial industry interested in large volumes and in few transactions.

Nevertheless, the benefits of crowdfunding are not just limited to raising money. Indeed, in accordance with the essence of crowdsourcing, backers often provide valuable feedbacks and ideas that fundraisers (i.e., those who publish projects and collect money through crowdfunding) can exploit to develop further their projects and entrepreneurial ideas (e.g., Colombo et al., 2015b; Riedl, 2013; Skirnevskiy et al., 2017). Specifically, crowdsourcing implies: the active engagement of the individual-backer, the emotional factor that represents an important driver of the funder's decisions and the high level of transparency, that is a crucial aspect on which the project initiators can leverage on to persuade the investors to participate and donate money.

2.2.2. Crowdfunding stakeholders

Crowdfunding platforms can be defined as multi-sided ones (Hagiu, 2006; Hagiu & Wright, 2015), since they act as intermediaries between money-seeking individuals and (potential) investors, thus connecting groups of customers in need of each other (Evans, 2003). The literature on crowdfunding has analysed three main categories of stakeholders: the *fundraisers*, *the backers,* and *the managers of crowdfunding platforms.*

Through online crowdfunding platforms, *fundraisers*, namely those who propose the ideas and/or projects to be funded (Mollick, 2014), meet potential funders (i.e., *backers*) that are searching for interesting projects and promising investment opportunities (Koch et al., 2018). Fundraisers can be defined as the entrepreneurial side of the platform, since they are the ones that initiate a campaign proposing an innovative idea, and in order to attract backers, they need to present their project in a very appealing way, using textual information, pictures or videos. Moreover, both a funding period (e.g., 30 days) and a funding goal (e.g., USD 20,000) have to be defined initially by the project founders (Koch et al., 2018).

On the opposite side of the platform, there are the *backers*, namely the ones that provide the financial resources to the fundraisers' projects. These investors play a fundamental role in providing feedbacks to project initiators and propose new ideas or modifications of the original project, valuable to define better versions of the initial concept (product/service).

Lastly, the *crowdfunding platforms*, which are intermediaries that enable transactions between fundraisers and backers, and they aim to reduce information asymmetries and thus the risks involved for the participating parties (Burtch et al., 2013). They own and manage the websites, serve as matchmakers between the sides' needs, and earn a fee or share on the transactions (Burkett, 2011; Koch & Cheng, 2016). Specifically, intermediaries play a very important role on equity-based and lending-based crowdfunding platform, because the campaigns and the requests for pledges are examined before their publication to confirm their reliability and lawfulness, in order to reduce risks for backers and to be in line with governmental policies.

2.2.3. Crowdfunding models

Modern crowdfunding can be represented by four distinct models. Starting from the definition of Lambert and Schwienbacher (2010), it is possible to identify different typologies of crowdfunding, on the basis of what funders obtain in exchange of their contributions. The different existent models are *reward-based*, namely investment in exchange for gifts or products, *equity-based,* i.e., investment for a percentage stake, *lending-based*, a peer-to-peer lending and *donation-based*, a charitable giving (J. A Boch et al., 2014). Reward-based crowdfunding is represented on well-known platforms such as Kickstarter and Indiegogo. The emerging model of equity-based crowdfunding is found on platforms such as Seedrs and Crowdcube. Donation-based crowdfunding, essentially known as charitable giving, can be found on many websites such as Just Giving, and a more prominent lending-based crowdfunding platform is The Funding Circle.

Crowdfunding Model	***Definition***
Reward - based	Individuals invest a predefined amount of money with the expectation that if successfully funded, they will receive a tangible (but nonfinancial) reward, product, or service.
Donation - based	Donation toward a specific project with no expectation of financial or material returns.
Equity - based	Small investments in crowdfunding project in return for an incremental stock in the respective business.
Lending - based	In the lending – based model, the crowd funder lends a small amount of money to a specific platform, project, or person.

Table 1: The four-crowdfunding model. Source Zhang et al., 2014 and Cholakova and Clarysse, 2015

Reward-based

A reward-based platform is where individuals fund a project or a new business in exchange of a non-financial reward, typically goods or services at a later stage. Typically, the creator of the campaign offers a unique service (reward) or a new product (pre-selling) in return for investment. This reward-based model is the most common for a number of platforms, such as Indiegogo, Fundable and Kickstarter.com, one of the oldest, largest, and popular crowdfunding platforms (Kuppuswamy and Bayus 2014). Usually, the platform gives two or more choices of contribution sorted by entity, with a different reward associated. In addition, in many cases, a reward crowdfunding campaign represents a way for gathering information about the demand for a specific product. For these reasons, usually, the price of the product/service offered in reward-based campaigns is lower than the price at which it will then be actually sold on the market. Depending on whether the funding target has been reached or not, such platforms follow one of these two schemes: *Keep-it-all* or *All-or-nothing*. This form of crowdfunding allows companies to launch with orders already on the books and cash-flow secured (a major issue for new business) and gathers an audience before a product launch (European Commission).

Kickstarter[7]

Founded in 2009, Kickstarter is one of the world’s biggest online reward-based crowdfunding platforms. The company's stated mission is to *"help bring creative projects to life"*. As of July 2021, Kickstarter has received nearly $6 billion in pledges from 20 million backers to fund 205,000 projects, such as films, music, stage shows, comics, journalism, video games, technology, publishing, and food-related projects.

Investors who backed Kickstarter projects are offered tangible rewards or experiences in exchange for their pledges. Indeed, Kickstarter requires creators to offer some kind of reward to their backers, no matter how simple or elaborate. When people fund a

[7] https://www.kickstarter.com/

project, they choose one of the predetermined awards the creators present. All Kickstarter pages have an *Estimated Delivery Date* section to specify when backers will receive their rewards. It may take several months before anything is delivered, especially if the reward is the product itself. On Kickstarter the *"all or nothing"* rule applies. Basically, a creator can collect funds only if the total amount of investments reaches the funding goal by the deadline, and Kickstarter put this rule in place to minimize risk for backers.

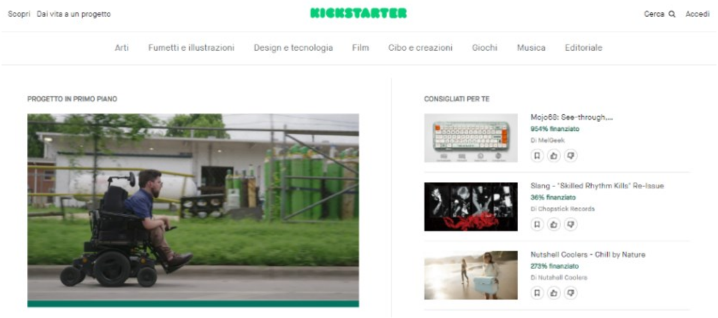


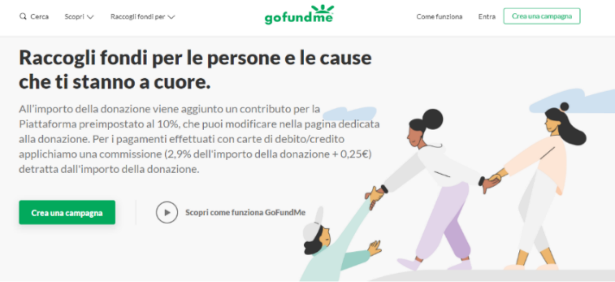
Figure 1: Kickstarter Webpage

Donation-based

Running alongside reward-based crowdfunding, donation-based is another very popular form of crowdsourcing. Donation-based crowdfunding can be defined as the collective effort of individuals to help charitable causes; indeed, funds are raised for religious, environmental, or other social purposes. he major aspect of donor-based crowdfunding is that there is no reward for donating; rather, it is based on the donor's altruistic reasoning. Donors come together to create an online community around a common cause to help fund services and programs to combat a variety of issues including healthcare and community development. Although there are no tangible returns for those who participate in such campaigns, very often the project's proponents tend to thank donors with intangible rewards, such as a thank you message or an image of the project realized.

GoFundMe[8]

GoFundMe is a free donation-based crowdfunding platform that allows creators to publish fundraising campaigns of various kinds, aimed at individuals, groups and organizations. GoFundMe enables, through the crowdfunding platform, to raise funds for various types of events or projects, which could be start-ups, beneficial events, celebrations and difficult circumstances. GoFundMe could be also worthy to raise funds for consequences of accidents and diseases. The platform is very active in fundraising, in 10 years from 2010 to 2020 were raised over 9 billion dollars, with over 70 million donors. Moreover, in many countries of the world GoFundMe is paid, but in Italy, United States, Canada, United Kingdom, Australia, France, Holland and Spain is completely free.



Raccogli fondi per le persone e le cause che ti stanno a cuore.

All'importo della donazione viene aggiunto un contributo per la Piattaforma preimpostato al 10%, che puoi modificare nella pagina dedicata alla donazione. Per i pagamenti effettuati con carte di debito/credito applichiamo una commissione (2,9% dell'importo della donazione + 0,25€) detratta dall'importo della donazione.

Figure 2: Gofundme Webpage

Equity-based

Equity-based crowdfunding, is a model in which backers provide money in exchange for a share of the risk capital of a firm. In this case, investors acquire ownership and voting rights, with the intention of participating in the distribution of future profits. Therefore, the investor acts as a shareholder who makes an investment and purchases

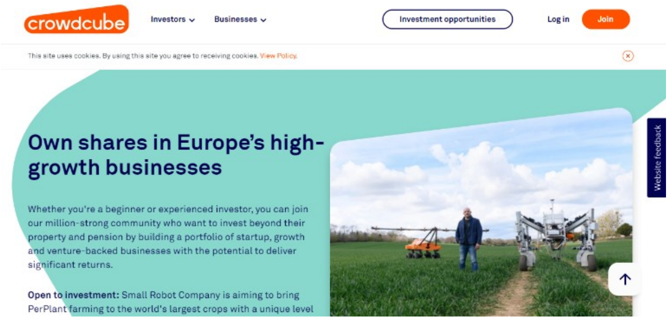
[8] https://www.gofundme.com/?utm_source=bing&utm_medium=cpc&utm_campaign=SE_GoFundMe_IT_EN_Exact&utm_content=Gofundme&utm_term=gofundme_e_c_&msclkid=5167ff06d2b71a81c4939e3ff90cad20

a share of the company property, in order to obtain dividends from the capital owned and possibly capital gains from the sale of his shares.

This crowdfunding model is highly regulated due to a high-risk profile and huge liabilities (Lasrado and Lugmayr, 2013). Moreover, equity crowdfunding, unlike donation and rewards-based, involves the offer of securities which include the potentiality for a return on investment.

CrowdCube[9]

Crowdcube is an equity-based crowdfunding platform founded by Darren Westlake and Luke Lang in 2011. With £689,234,678 collected and a community of over 750,000 users, over 880 campaigns have been funded on Crowdcube. The key principle of the platform model is based on the fact that anyone can invest money in an asset, in exchange for an equity return. Entrepreneurs with a UK-registered business can show their entrepreneurial projects to thousands of potential micro-investors by uploading videos, images and documentation. The minimum investment is £10. Crowdcube operates with a policy *"all or nothing":* when the target is reached, the company receives the funds, otherwise everything is returned to investors. A commission is paid if the fundraising campaign has been successful.



This site uses cookies. By using this site you agree to receiving cookies. View Policy.

Own shares in Europe's high-growth businesses

Whether you're a beginner or experienced investor, you can join our million-strong community who want to invest beyond their property and pension by building a portfolio of startup, growth and venture-backed businesses with the potential to deliver significant returns.

Open to investment: Small Robot Company is aiming to bring PerPlant farming to the world's largest crops with a unique level

Figure 3: Crowdcube Webpage

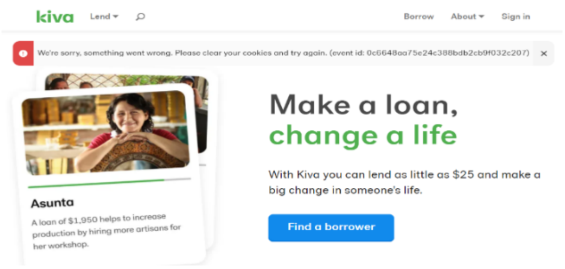
[9] https://www.crowdcube.com/

Lending-Based

Lending-based crowdfunding (also called *"peer-to-peer lending"*) means that businesses can borrow directly from tens or hundreds of people who are available to lend. It could be considered a different option respect to a bank loan, and instead of borrowing from a single source, the crowd provides the investment. Internet-based platforms are used to match lenders with borrowers: the first ones, who need a loan in order to develop a project, make the request to the online platform, on which the lenders invest by providing money at an average interest rate higher than those offered by credit institutions.

Kiva[10]

Kiva is the world's most popular lending-based crowdfunding platform based on the mission to alleviate poverty by connecting people through lending. Through crowdfunding, early-stage enterprises are able to tap a broader network of smaller, individual lenders that they otherwise would not have been able to access. Kiva, indeed is a global *"network"* that, through its online portal, connects individual financiers and small entrepreneurs in developing countries. This is concretely possible thanks to the partnership with the microcredit institutes (*Field Partners*) that, from Kenya to Ecuador and all over the world, connect to *kiva.org* and publish the photos and profiles of small entrepreneurs in need of loans, after selecting their best projects short of funds.



We're sorry, something went wrong. Please clear your cookies and try again. (event id: 0c6648aa75e24c388bdb2cb9f032c207)

Make a loan, change a life

With Kiva you can lend as little as $25 and make a big change in someone's life.

Asunta

A loan of $1,950 helps to increase production by hiring more artisans for her workshop.

Figure 4: Kiva Webpage

[10] http://www.kiva.org/

2.2.4. Crowdfunding elements

The final aim of each crowdfunding campaigns' proponent is to raise funds in order to finance the innovative idea: the first step is to draw the attention of backers on the proposed project. To this end, fundraisers can leverage five different campaigns' features that will be properly explained in the following section: (1) *project description*, (2) *target capital*, (3) *duration of the campaign*, (4) *rewards* (when present), and (5) *fundraisers' information*. (Butticè et al., 2018)

Project description

Presenting the project in an attractive way to backers, is one of the first levers entrepreneurs can adopt in order to raise their funds. Usually, the main information the proponent has to insert in the campaign are the *title*, that has to be very appealing and summarize the main concept of the project, a *blurb*, in which the project is summarized, and a detailed *project description* in which all the characteristics of the initiative are presented to potential investors.

The detailed project description is the main source through which fundraisers drive information to the backers and could become a fundamental source of engagement within the crowd, particularly when the textual description is accompanied by a video pitch and/or images showing the characteristics of the product/service proposed. The availability of rich information has important implications for potential contributors' decision-making. Moreover, during the campaign the fundraiser has the possibility to furtherly inform the crowd about new developments of the project through *updates*, which are published on a voluntary basis. *Comments* left by platforms' users are another source able to attract new backers, and they are employed as an instrument to ask for information or to cheer and thank the fundraisers for their products/services. Then, the fundraisers can reply directly to the comments posted.

Target capital

The target is one among the crucial elements provided and set by the creator of the project and not editable once the campaign has started. Clearly, for fundraisers it is fundamental to define the most appropriate value of the target-sum as only if the latter is reached, then the campaign will have been successful. Several factors influence which is the best amount to fix, such as knowing how large the reachable crowd is and how it is possible to get in contact with investors. Specifically, two different fundraising schemas exist in crowdfunding platforms, namely *"keep-it-all"*, according to which the creator will obtain the money collected regardless of the target reaching and *"all-or-nothing"*, which delivers the money obtained through backers' investments to the creator only if the target-sum is achieved.

Fundraisers can set a target and in a second moment, if the pledges exceed the initial goal, can encourage a further crowd participation through stretch goals, promptly communicated to potential investors through *updates*. Also, there is the possibility that the total amount of backers' contributions exceeds the target fixed by the creator, and this phenomenon is called "*overfunding*" (Li Y. et al., 2020).

Duration of the campaign

The duration of a campaign is another of the elements defined by the fundraiser, and it should be in line with the policies provided by the platform where the project is launched. For example, considering *Kickstarter.com*, campaigns can last anywhere from 1 to 60 days. Therefore, the duration decided by the creator cannot exceed a maximum of 60 days. According to the support of the Kickstarter website[11], the following sentences are reported: "*We've done some research, and found that projects lasting any longer are rarely successful. We recommend setting your campaign at 30 days or less. Campaigns with shorter durations have higher success rates, and create*

[11] https://help.kickstarter.com/hc/en-us

a helpful sense of urgency around your project." For this reason, often fundraisers prefer a 30-day duration for their campaigns.

The achievement of the target has to happen within the timeframe of the campaign, otherwise the campaign will be considered unsuccessful (Colombo et al., 2015b).

Rewards

As mentioned in the previous section, different crowdfunding models entail the provision of a reward to investors, whose value is proportional to the amount of the money pledged by the backers.

The literature distinguishes between two types of reward (Boeuf et al., 2014): symbolic rewards and material rewards. The former consists in acknowledging the backers for the support received through intangible rewards, both privately such as a thank-you email, or publicly, for instance by listing the bakers' name on a website or in the credits of a movie (Butticè & Colombo, 2017). Instead, the latter consist in tangible rewards such as gadgets and gifts, or in a product, which is often the outcome of the project for which funding is sought. Moreover, since it is up to the fundraiser to deliver the promised rewards to the supporters at the end of the campaign, the credibility of serial fundraisers (i.e., those who launch several campaigns on crowdfunding platforms) is also dependent on their previous track record and reliability.

Fundraisers' information

As reported at the beginning of the Chapter (*2.2.1*), Venture Capitals and other financial intermediaries' funds often follow criteria to select investment targets which are more restrictive than crowdfunding criteria. Indeed, in the most traditional financing sources, investors require to analyse toughly the person who is asking for money, to assess the degree of risk associated with the lending activity. Instead, in the crowdfunding platform, the borrowers involved are often people with a less developed knowledge on credit risks and liabilities, and so the assessment developed by potential backers is also less structured than the one applied by financial actors.

Since the fundraisers' information such as a personal description and photo have to be disclosed and posted on a voluntary basis by the fundraisers themselves, there may be a heterogeneity both in quantity and quality of the available data. However, potential backers can find information about the human and the social capital of the fundraiser through the profile posted on the platform, or alternatively, they can identify previous campaigns initiated on the same crowdfunding platform, and access the previous track of record and reliability of the proponent.

2.3. Crowdfunding Success

As previously anticipated, before starting the study on the core theme of the project, namely overfunding, the following chapter is dedicated to provide an overview on the concept of success of a crowdfunding campaign and its main determinants, which are organized according to the following classification: *campaign-related factors* (1), *fundraiser-related factors* (2), *backers-related factors* (3) (Kaartemo, 2017).

We believe that a careful study of the dynamics of success is a fundamental preliminary step to be able to point out the peculiar theme of having *"more than success",* and above all, discriminate between success and overfunding.

2.3.1 Success: an overview

The enraptured growth of the crowdfunding sector has attracted the attention of many scholars over the years, who have directed their studies on successful campaigns and all the elements affecting the possible positive outcome of a crowdfunded project.

As previously mentioned, at the very beginning of a campaign, fundraisers have to determine the length of the funding period and to fix the goal to be reached. One main metric, to measure crowdfunding success, is meeting the target amount within the project duration, i.e., raising at least the amount of money stated as the campaign goal within the time span of the campaign (Colombo et al., 2015b). However, it is possible to say that the latter definition applies only to platforms that follow an *"all or nothing"* model in which, if the target is not reached, the money is returned to investors.

More in general, according to Butticè et al., the basic definition of success is clarified as reaching the target capital within the campaign duration. (e.g., Barbi & Bigelli, 2017; Koch & Siering, 2015; Mollick, 2014; Zvilichovsky et al., 2013).

A very interesting point is that, regardless of its last-years popularity, statistics show that the vast majority of crowdfunding campaigns dramatically fail with 81% of failed campaigns reaching less than 20% of their funding goal (Forbes et al., 2017). Of course, there are many reasons and factors that can influence the success or failure of a campaign, and according to this, more and more scholars in the last years have focused their research on the factors that drive the outcome of crowdfunding campaigns.

2.3.1. Main success factors

As reported in the previous statement, many papers deal with the factors that determine the success of crowdfunding campaigns. One of the most evident examples is the research of M. Barbi and M. Bigelli about the crowdfunding practices in and outside the US, which states that a more detailed description of the project increases the probability of funding. Indeed, an accurate description could be interpreted as an additional quality signal of the project, as long as it is not too extended because this could be harmful. Moreover, according to other scholars, a higher probability of success is significantly affected by: the presence of a good video presentation, a higher number of reward levels, a relatively short campaign duration and a smaller funding target, the reputation and reliability of those who present the project (Wang et al., 2018; Kuppuswamy and Bayus, 2015; Li and Martin, 2014; Mollick, 2014). In the last few years, a new trend has been developed: researches are not only interested in the factors determining the success of a campaign, but also in the effect of their interaction, such as in the 2019 research of J.A. Koch, *"The recipe of successful crowdfunding campaigns - An analysis of crowdfunding success factors and their interrelations"*, whose empirical results, which are also based on previous studies on the theory of signals and investment decision making, reveal that not only the factors themselves, but especially their interrelations, may promote the success of the financing.

In particular, among all the factors studied in the literature, it is also important to point out that in recent years researchers have not only focused on "*static*" factors, but also on the impact that *dynamic* variables can lead to the outcome of the project. The main difference between static and dynamic determinants is that the formers are collected in the moment a campaign is launched, and they remain fixed until the end of the campaign, while the latter factors are instead those that change as the crowdfunding campaign progresses.

In the next paragraphs, a study of all the most important studies is reported, concerning static factors divided in *campaign-related*, *fundraisers-related*, *backers-related* (Butticè et al., 2018), then linguistic features and dynamic ones.

2.3.1.1. Campaign-related factors

In the following section, we discuss the main campaign-related features affecting campaigns' success. The features presented and analysed are the *typology of project content*, the *funding target*, the *campaign duration*, the *number and typology of rewards levels*, the information made available to potential backers, such as the presence of *videos* and *images*, *comments* and *updates*.

Starting with the *project content*, crowdfunding typology of campaigns varies across a wide array of categories, and depending on these sectors they are more or less likely to be successful. For example, Colombo and colleagues (2020) underline that many projects are aimed at developing products with a technological core (Colombo et al., 2015; Mollick, 2016), and the probability of being funded varies depending on the type of innovation proposed (*incremental* or *radical*). In particular, Chan and Parhankangas (2017) state that projects implying incremental innovation are more likely to achieve success, as they are not only more feasible, but also involve less learning effort and risk for backers compared to projects characterized by radical innovation, which, on the contrary, tend to receive less funding. Aside from technological projects, many campaigns have an orientation to sustainability, and this has an impact on success as well. Calic and Mosakowski (2016) found that these projects have a higher chance of receiving funds. Another important criterion of distinction that has proved useful is whether the project is non-profit versus for profit, the former being found more likely to succeed than the latter. (Liao et al., 2015; Pitschner & Pitschner- Finn, 2014).

Setting the right *target* (i.e., the amount the fundraisers seek to raise using crowdfunding (Mollick, 2014)) at the beginning of the campaign is unanimously considered by scholars as a crucial determinant of success. Indeed, numerous studies found consistent evidence of a negative relation between the target capital and a project's success, precisely, the higher the target capital the lower the probability of success (see e.g., Colombo et al., 2015b; Gleasure & Feller, 2014; Liao et al., 2015; Mollick, 2014; Zheng et al., 2014, among others). Among them, Colombo et al. (2015b) have noticed that backers increase with the target capital, but in this case, they provide smaller amounts of money. In addition, Frydrych et al. (2014) find that projects with a higher target capital experience great difficulty in achieving legitimacy and their fundraisers have to make more effort to obtain funds.

Regarding the effect of the *campaign duration* on the probability of success, the various studies on the subject have led to conflicting results. Some scholars (Frydrych et al., 2014; Mollick, 2014) underline a negative impact of the campaign duration on the success, since a longer funding period could be seen as a signal of a lack of confidence by investors. A completely different point of view is the one of Liao et al. (2015), affirming that longer campaigns favour the achievement of the target capital. The conflicting views may depend on the market under analysis, namely the USA for

Frydrych and colleagues (2014) and Mollick (2014) and China for Liao and colleagues (2015) (Butticè et al., 2018).

The role of *rewards* in determining success has to be analysed according to two dimensions: the number and the types of rewards (Boeuf et al, 2014; Gerber & Hui, 2013). Regarding the first factor, consistent evidences exist that when a project provides a wide variety of rewards among which backers can choose from, the probability of success is higher (Kunz et al., 2017). Consistently, Mollick and Nanda (2016) suggest that a greater number of reward levels is associated with good crowdfunding performance. Nonetheless, the number of reward tiers should not necessarily be extended indefinitely, as fewer meaningful reward tiers are characteristic to successful campaigns (Chen et al., 2016), since also the rewards' quality has to be taken into account (Hörisch (2015); Hobbs et al. (2016)).

The last campaign-related factor influencing the probability of success is the *quality* and *amount of information disclosure* about the campaign itself, since it is important to signal the campaign quality through communication (Kaartemo (2016)). Numerous scholars (Fondevila Gascon et al. (2015); Hobbs et al. (2016); Mollick (2014); Mollick and Nanda (2016)), show that signals of quality including campaign *video*, *pictures*, content precedence, detailed *text description*, rapid *updates*, proofread, increase the likelihood of meeting the funding targets.

Campaign-related factors	***Effect on campaign success***	***Source***
Category	There is not a specific category able to attract more funds.	• Dushnitsky and Fitza, 2018
Type of innovation	Project proposing incremental innovation are more likely to succeed than the ones proposing radical innovation.	• Chan and Parhankangas, 2017

Sustainability – oriented projects	Projects with a sustainable aim are more likely to be funded.	• Calic and Mosakowsi, 2016
Non-profit versus for profit	Non-profit campaigns are more likely to be funded instead of for profit.	• Pitschner and Pitschner-Finn, 2014 • Liao et al., 2015
Funding target	The higher the target capital, the lower the probability of success	• Frydrych et al., 2014 • Koch and Siering, 2019 • Colombo et al., 2015b • Gleasure and Feller, 2014 • Liao et al., 2015 • Mollick, 2014 • Zheng et al., 2014 • Shenor and Vikk, 2020
Campaign duration	Diversified results	• Scholars underlying a negative impact of campaign duration on success: • Frydrych et al, 2014 • Mollick, 2014 • Scholars underlying a positive impact of a longer campaign duration on success: • Liao, 2015

		- Lagazio and Querci, 2018 - Shneor and Vikk, 2020
Rewards	Projects offering a wide variety of rewards to backers have higher success probability. Nonetheless, the number of rewards should not be extended indefinitely. The higher the quality, the higher the probability of success	- Kunz et al., 2017 - Mollick and Nanda, 2016 - Chen et al., 2016 - Hobbs et al., 2016
Quality and amount of information made available	Signals of quality including campaign video, pictures, content precedence, detailed text description, rapid updates, proofread, increase the success probability.	- Fondevila Gascon et al., 2015 - Hobbs et al., 2016 - Mollick, 2014 - Mollick and Nanda, 2016

Table 2: Impact of campaign-related factors on success

2.3.1.2. Fundraisers-related factors

The main elements associated to the fundraiser's characteristics, which drive the result of a campaign, have aroused much interest in scholars' point of view. Among these factors, the most cited in the literature are *gender*, *race* and the *team composition* of the fundraisers. Subsequently, it is also worthy to mention the fundraiser's *human and social capital*.

Identity information reveals that identity disclosure (by showing a name and photo) strongly affects how supporters perceive trust. Similarly, in the crowdfunding context, a project with the clear identity of the founder will gain more credibility, which in turn leads to a higher potential for reaching the fundraising goals (Kim et al., 2017). More in depth, when a fund-seeker is an individual, *gender* might have an influence. There are several studies demonstrating that project created by female entrepreneurs, on reward crowdfunding platforms, experience higher success rate than male-created projects (Colombo et al., 2015b; Frydrych et al., 2014; Greenberg and Mollick, 2017).

The role of an individual or an organization that asks for money also influences crowdfunding performance. For this reason, a noteworthy aspect to consider when investigating crowdsourcing determinants of success is the role of *team versus individual fundraisers*. These differences are studied more at an organizational level. Fund-seekers that have B2C projects are more likely to succeed than those with B2B projects (Lukkarinen et al., 2016). Moreover, those projects featuring pairs or teams demonstrate higher success rates than projects carried out by individual fundraisers. Likely, this effect is due to the fact that a team might show a more diversified and skilled pool of resources and it might also rely on a wider network of contacts (Lagazio and Querci, 2018).

Following the definition of Davidsson and Honig (2003), *human capital* is an intangible asset, possessed by a team or an individual, representing both *tacit knowledge*, gained through experience, and *explicit knowledge,* gained through formal education. In the context of crowdfunding, human capital can be a key signal to evaluate projects (Yeh and Chen, 2020) and it can effectively impact on the likelihood of funding success (Ahlers et al., 2015). A wide stream of the extant literature addressed the role of this factor in affecting the success dynamics (Butticè et al., 2017; Skirnevskiy et al., 2017; among others), finding that prior experience in crowdfunding is a way for easing the fundings collection. These results suggest that previous activities might signal higher reliability compared to fundraisers who have not been active before (Koch and Siering, 2019).

The *social capital* of the fundraiser can originate from multiple sources: as noticed by Colombo et al. (2015b), Liao et al. (2015) and Cai et al. (2020), two types of social capital exist, namely the internal and external social capital. The former refers to the social network created within the crowdfunding platform; the latter represents the social capital developed outside the platform (Cai et al., 2020). With respect to internal capital, Butticè and Useche (2019) suggest that developing a robust network of connections within the platform, entrepreneurs are able to overcome the liability of *"outsider-ship"*, which is a phenomenon that characterizes traditional forms of financing and that stems from the lack of local connections. While Liao et al. (2015) contend that both types of social capital (i.e., internal and external) have a positive effect on crowdfunding success, Colombo and colleagues (2015b) state that only internal social capital has a relevant effect on success, specifically it helps in obtaining more funds in the earlier stages of the campaign. Instead, Cai and colleagues (2020) claim that the relevance of internal and external social capital changes over time. During the earlier stages of the campaign, since most of the funds come from the direct relationships of the fund seeker, the development of external social capital is essential to effectively promote the campaign also among unknown people. As the campaign evolves, instead, internal social capital reveals to be more significant than the external one, due to the *word-of-mouth* communication among other potential backers.

Fundraisers-related factors	***Effect on campaign success***	***Source***
Gender	Project created by female fundraisers experience higher success rate than male-created projects	• Colombo et al., 2015b • Frydrych et al, 2014 • Greenberg and Mollick, 2017 • Moleskis et al., 2019 • Zhao et al., 2020
Team versus individual fundraisers	Projects featuring pairs or teams demonstrate higher success rates than projects carried out by individual fundraisers.	• Frydrych et al., 2014 • Lagazio and Querci, 2018

		• Horvat and Papamarkou, 2018
Human Capital	Having prior experience in crowdfunding eases the funding collection	• Butticè et al., 2017 • Skirnevskiy et al., 2017 • Cappa et al., 2020
Social Capital	Mixed results	Scholars claiming only internal social capital matters and has a positive impact on success: • Butticè and Useche, 2019 • Colombo et al., 2015b Scholars contending both internal and external social capital have a positive impact on success: • Liao et al., 2015 • Cai et al., 2020

Table 3: Impact of fundraisers-related factors on success

2.3.1.3. Bakers-related factors

Backers play a crucial role in crowdfunding, since they are the ones providing funds to projects thus determining their success. Previous researches have shown that some investment measures, such as capital raised over time, follow a *U-Shaped* trend along the duration of a campaign. Indeed, projects receive much more funds at the beginning and at the end of their duration, while investments are considerably lower in the middle of the campaign.

Several studies have analysed backers' motivations and characteristics, finding a correlation between them and the success of crowdfunding campaigns. There are three main elements to be taken into consideration: *strategic motives, social identification* and *group identity.*

With respect to the first one, Berns and colleagues (2020) and Nielsen and Binder (2020) contend that a fundraiser is able to attract more backers when the *strategic motives* associated to the campaign are presented, instead of the altruistic ones.

As per the *social identification*, the extant literature suggests that successful crowdfunding projects match with what is valued by the crowd funders (Zheng et al., 2014). Indeed, in order to be effective, the campaign should be designed in a way that allows a social identification from the crowd (Martens et al., 2007). To do that, it is necessary that the values and principles featured in a project are aligned with those of potential investors (Nielsen and Binder, 2020).

The last investigated factor is the *group identity*. Belleflamme et al. (2014) encouraged crowdfunding fundraisers to build a community to make the backers feel involved in the project they are financing, thus becoming more willing to contribute. With the final aim of creating a community and a group identity, entrepreneurs can leverage on crowdfunding platforms' comment section to foster their online interactions.

Bakers-related factors	***Effect on campaign success***	***Source***
Strategic motives	A fundraiser is able to attract more backers when the strategic motives associated to the campaign are	• Berns et al., 2020 • Nielsen and Binder, 2020

	presented, instead of the altruistic ones.	
Social identification	Projects whose value and principles are aligned with those of potential investors are more likely to be funded.	• Nielsen and Binder, 2020
Group identity	Fundraisers able to create a community are going to benefit from higher contributions.	• Belleflamme et al., 2014

Table 4: Impact of backers-related factors on success

2.3.1.4. Linguistic factors

Linguistic features play a crucial role in the communication process, for example the narrative pitches are strongly considered by potential backers while evaluating a project. Such cues have varied effects in stimulating the backers' capital provision to a specific project (Yuan and Wang, 2020), thus affecting the success rate of the fundraising.

The relationship between the linguistic features of a crowdfunding campaign and its success, in recent years, is becoming a topic of interest in literature. In the following section we will classify all the language features of a campaign, in particular those used in *project description*, *title* and *blurb*, and those used in other parts of the campaign, such as in the *comments* section and in the *updates* section. The classification was derived from a careful reading of the analysis carried out by Annunziata and Aversa, 2020.

Linguistic factors	***Effect on campaign success***	***Source***
Money-related language	Terms related to money are negative predictors of crowdfunding success, except in the case of civic projects.	• Allison et al., 2015 • Chan et al., 2019 • Kaminski and Hopp, 2020 • Xu, 2018 • Lee et al., 2019
Risk-related language	There is an inverted u-shape relationship between the signals of risk taking and crowdfunding success.	• Allison et al., 2015 • Calic and Shevchenko, 2020 • Lee et al., 2019
Prosocial words and altruistically framed campaigns	The number of prosocial words used is a significant predictor of campaign success (except in comics, dance and theatre categories). Consistently, altruistically framed campaigns have more probability of being funded, compared to campaigns that focus on egoistic motives.	• Pietraszkiewicz et al., 2017 • Nielsen and Binder, 2020
Green words	Mixed results.	No effect on probability of success: • Hörish, 2014 Positive effect on the probability of success for technology projects:

		• Calic and Mosakowski, 2016
Building relationship and gratitude-related cues	They are important success predictors for community and creative projects	• Yuan and Wang, 2020
Reward-related cues	They positively affect technological and innovation campaigns.	• Yuan and Wang, 2020
Need-driven narrative	There is a negative relation between the use of the words “help” and “thank” and the probability of success. Indeed, scholars contend that a need driven narrative is detrimental to the amounts of funds received.	• Xu, 2018 • Berns et al., 2020
Narcissistic language	A moderate amount of narcissistic rhetoric can have a positive impact on crowdfunding success. However, an excessive level can become detrimental, since it may convey instability and untrustworthiness. Narcissism is very penalized in industries in which the value of the product is strongly associated to that of the fundraiser, such as art, design, film, food, journalism and theatre.	• Anglin et al., 2018 • Butticè and Rovelli, 2020
Perceptual words	The use of positive affective and perceptual language when presenting the project pitches can lead to a greater likelihood of success.	• Lee et al., 2019
Positive tone	Project success increases proportionally with the usage of positive tone, as people employing a positive emotional language are	• Gorbatai and Nelson, 2015 • Wang et al., 2017

	more likely to be perceived optimistic and confident.	• Zhou et al., 2018
Inclusive language	It is a positive predictor of a campaign success, since an inclusive language make investors feel connected to the cause or the funder.	• Kaminski and Hopp, 2020 • Gorbatai and Nelson, 2015
Concrete language	Concreteness refers to a linguistic style representing contextualized and detailed representations of objects (Doest et al., 2002), and it is positively related to success in case of social campaign, while it has no impact on commercial campaign success.	• Parhankangas and Renko, 2017 • Koh et al., 2020
Result in progress narrative (RIP) and Ongoing journey (OJ) narrative	Two linguistic narratives have been investigated by scholars: Result in progress (RIP), connected to product features, and Ongoing journey (OJ), emphasizing intangible aspects like values, principles and visions during an entrepreneurial project. Results on the connection between the twos and success contend that fundraisers with little experience receive more pledges by adopting RIP narratives, while experienced entrepreneurs obtain more funds using OJ narratives.	• Cappa et al., 2020
Agentic and communal language	A relationship between gender and the language of crowdfunding has been investigated. The likelihood of campaign' success decreases if a female fundraiser adopts an agentic language.	• Butticè and Rossi-Lamastra, 2020

Table 5: Impact of linguistic features adopted in the detailed project description on success

Linguistic factors	***Effect on campaign success***	***Source***
Positive language	Using positive sentiment in the blurb promotes campaign success, while its use is discouraged in the title since it results to be negatively correlated with the campaign success.	• Wang et al., 2017
Social language	Social framing has a negative effect when present in the title and/or in the blurb.	• Defazio et al., 2020

Table 6: Impact of linguistic features adopted in title and blurb on success

Lastly it is interesting to focus on the use of linguistic features in the comment section and in the updates one, since they represent a powerful leverage that can have a strong impact on campaign's success.

Linguistic factors	***Effect on campaign success***	***Source***
Positive language	Using positive tone and sentiment promotes the success of a crowdfunding campaign.	• Jiang et al., 2020 • Wang et al., 2017
Simple language	Updates that use a simple language have a positive impact in the short term and are useful during the central phase of the campaigns	• Block et al., 2018 • Ryoba et al., 2020

Table 7: Impact of linguistic features in comments and updates on success

2.3.1.5. Dynamics affecting crowdfunding success

As reported at the beginning of the Chapter (*2.3.*), dynamic factors are the one changing as the crowdfunding campaigns progresses. It is trustworthy, according to the papers of the literature, that there are not only static factors (fixed before the campaign begins) to define the success of the project itself. Rather, some elements, the dynamic ones, can be modified during the life-cycle of the project campaign, so as to attract and convince more investors to finance.

Chen et al. (2020) introduced the concept of life cycle stages associated with the campaigns, and they demonstrated that the determinants of success vary in three different stages (growth, stable and maturity stage). Indeed, this happens because at the time of the launch of a campaign, backers are able to rely only on static determinants since dynamic information has not been produced yet. Then, with the progress of the project, dynamic features become increasingly available and play an incrementally important role for funders. Specifically, in the first phase, it is likely that there are several and rapid contributions, mostly coming from friends and family of the fundraisers (Agrawal et al., 2015); the second phase, instead, is considered the most difficult to overcome, since just few campaigns are able to go beyond it. Indeed, in this step, monetary contributions tend to reduce (Ordanini and Parasuraman, 2011). Finally, just for few of the projects that successfully overstep the previous phase, it is possible to assess a rapid growth of the contributions which eventually leads to the achievement of the funding goal (Crosetto and Regner, 2014; Kuppuswamy and Bayus, 2017; Ordanini and Parasuraman, 2011).

Typically, project creators rely on communication to make investors feel involved and finance the initiative. For example, in the last phase of the campaign duration, the growth in contributions can be explained through the phenomenon that pushes people to take action when they can feel their financial support is going to actually make a difference for the fundraiser, and this can be amplified through the use of the right communication-method. Communication is not only crucial in the last phase to push backers, but in each phase of the campaign lifespan: the findings of Ryoba and colleagues (2020) show that, at the first stage of the campaign life-cycle, communication should focus on the *quantity of updates*, since the more updates are delivered at the beginning of the campaign, the easier is for potential investors to make informed decision. Also, the *tone of the comments received* is an important parameter to assess if the project is going to be successful or not, as a positive tone is able to attract more backers, while negative a one is a strong indicator of difficulties in receiving future investments. In the second phase of the campaign, the *readability of the updates* becomes more relevant than their quantity, because backers are more interested in knowing how fundraisers intend to achieve the planned objectives. Finally, in the last phase of the campaign, the most important communication aspect in predicting success is the *sentiment polarity of the comments* received by the investors (i.e., their tone) and this shows that potential backers are affected in their funding decisions by the opinions of previous investors.

Even if the communication is crucial for the entire lifespan as reported in the previous paragraph, it is important to notice that the most insightful data on the possibility of the campaign to reach its goal are collected during the first days of its activity**.** Related to this, Colombo and colleagues (2015b) find that the *number of backers* and *the percentage of funding target* raised in the first sixth of the campaign's duration are positively associated to the its success. Deepening, Colombo and colleagues have confirmed the thesis supported by Petitjean (2018), who studied how the success-factors evolve during the campaign, revealing the importance of the performance of the first week. He finally found out that the percentage of the target funded after one week can be considered as an indicator of the probability of success. This finding is in line with other scholars, who state that previous contributions play an important role in determining the success or failure of a campaign.

Dynamic factors	*Effect on campaign success*	*Source*
Updates	The more updates are delivered at the beginning of the campaign, the higher the probability of reaching success. Moreover, the readability of updates is fundamental during the last phase of the campaign.	Ryoba et al., 2020
Tone of the comments	A positive tone is able to attract more bakers, thus increasing the probability of success.	Ryoba et al., 2020
Early-stage contributions	The number of backers and the percentage of funding target raised in the first sixth of the campaign's duration are positively associated to the campaign success.	Colombo et al., 2015b Vismara, 2016

Table 8: Impact of dynamic factors on success

2.4. Overfunding

Overfunding is a distinct feature of crowdfunding, whereby funders continue to invest in *"overly successful"* campaigns even beyond the goal stipulated by fundraisers (Li Y. et. al., 2020). As previously mentioned, despite the growing literature on crowdfunding in the last years (Mollick, Vulkan et al., Vismara, Agrawal et al., Cumming et al., among others), there is a paucity of studies which have examined the peculiar phenomenon of overfunding. Some examples are constituted by the findings of Koch, Cordova et al., Adamska-Mieruszewska et al. in reward-based fundraising, and Xiaoyu et al., Li et al. in equity crowdfunding.

In this section we present all the main topics discussed in literature, clustering them into the main areas of interest. Indeed, we carefully read and analysed the selected articles and papers (that we already discuss in *paragraph 2.1.*), and since they address several topics from different perspectives, we grouped them into three main categories, summarized below:

1. Overfunding as a distinctive feature of crowdfunding, which is an open innovation paradigm, different from traditional funding methods (Business Angels and Venture Capitalists): in the analysed papers emerged that the overfunding phenomenon does not exist in the case of traditional financing, in which the applicant gets the exact sum of money required;
2. Overfunding as a source of market inefficiencies. This paragraph is divided into three sub-paragraphs:
   - Overfunding as a cause of the *"market for lemon"* phenomenon, due to an inefficient distribution of resources, leading to few projects being over-financed and many quality projects not even reaching their target;
   - Overfunding with negative implications for the campaign creator, that in the majority of the cases may not be able to manage the excessive resources received. Some scholars, on the other way around, contend that overfunding may have a positive impact on the creator's reputation;
   - Overfunding as a source of negative implications for backers, focusing on the problem of adverse selection in crowdfunding markets;
3. Overfunding determinants, as the main factors leading to the phenomenon.

In the subsequent paragraphs we will deepen each of these topics addressed in recent years, in order to have a general view of which are the main dynamics of the phenomenon and its main determinants, especially to be able to identify which are the areas that have not yet been considered by scholars (*"gaps in the literature"*). The latter will be introduced in the *Research Objectives Chapter.*

2.4.1. Overfunding as a distinctive feature of crowdfunding

It is necessary to state that entrepreneurial initiatives need monetary resources to survive and grow (Block et al., 2018a). Specifically, raising initial capital is one of the main problems entrepreneurs have to face at the seeding stage of their business ventures. Even though multiple sources of financing like banking institutions, private foundations, governmental agencies, business angels (BAs), and venture capitalists (VCs) are available for fund raising, they are less accessible to novel entrepreneurs who lack all sorts of guarantees such as the evidence of innovative capability and venture profitability. Indeed, the above-mentioned investors are risk-adverse, thus they tend to disregard business ventures with high return variability, unproven profitability and whose value is difficult to be assessed (Li Y. et. al., 2020).

According to this, start-ups are not suitable to use the traditional funding channels because of the main characteristics of their structure: weak capability to produce income in the first years of life and therefore inability to pay interests and instalments, high risk inherent in innovative businesses, lack of capital guarantees and difficulties for banks to assess the real value of the company. According to *"Go globe, Startups success and failure rate"* published in *Statistics and Trends*, the 82 % of start-ups are self-funded or supported by the entrepreneurs' friends and relatives, whereas less than 1% of these entrepreneurs managed to raise initial capital from VCs.

For all these reasons, new ventures and entrepreneurial initiatives resort to innovative funding paradigms, such as the crowdfunding one, which is an open innovation model implying an open call, mostly through dedicated platforms on the internet for the provision of financial resources, either in form of donation or in exchange for future products or some form of reward (Belleflame et. al., 2014). Unlike traditional financing methods, where the applicant receives precisely the amount of money required, in the case of crowdfunding the pledge can also exceed the target within the campaign duration, thus causing the overfunding. This phenomenon is only possible in the case of funding through crowdfunding platforms, in which potential investors can participate in campaigns they consider worthy even beyond the target. Basically, relying on the *"crowd"*, project creators and start-up initiators can receive a higher amount of money much faster compared to traditional paradigms. Moreover, overfunding can be also highly beneficial by acting as an implicit certification of the firm's quality and sending a positive signal to potential investors such as venture capitalists and business angels (Coakley et al., 2018), thus solving the problems of the lack of guarantees for new ventures.

2.4.2. Overfunding as a source of market inefficiencies

Despite the overfunding, seen in absolute terms, could be considered as a highly positive phenomenon, the existing literature has underlined some negative effects that it causes on the market, arising from an inefficient resource allocation. Conceivably, overfunding can be deemed to be one of the primary causes of market inefficiency for crowdfunding platforms because of its detrimental effects on the long-term benefits of platform providers *(4.2.1),* campaign creators *(4.2.2)*, and backers *(4.3.3)* (D. Liu et. al., 2015).

2.4.2.1. Market for lemon

In crowdfunding markets, overfunding may become a primary source of market inefficiency since it promotes a suboptimal allocation of resources towards highly visible campaigns (Li et al., 2020). Indeed, overfunding can cause externalities with an impact on other projects on the platform. Here, both positive and negative external effects of overfunding are found and discussed (Doshi, 2014; Kim et al., 2016; Liu et al., 2015). The reason for such effects is that crowdfunding projects on a platform compete for funding (Burtch, 2011): money that is spent on one project cannot be received by another one.

These observations led Yen C.-H. and colleagues (2020) to speculate that the current *"invisible hand"* approach of matching massive numbers of small donations to crowdfunding projects is suboptimal, in the sense that the distribution of fundings could be improved so that more high-quality projects would succeed. This may result in some very good projects being unable to reach their donation goals, while a few superstar projects receive several times the amount of money they were initially seeking. Similarly, Martínez-Gómez C. and colleagues (2020) contend that sustainability of financial system requires an optimal allocation of scarce monetary resources among investment alternatives, and so overfunding becomes a source of inefficiency.

On the same line of thought are placed Li Y. and colleagues (2020), who contend that, by funnelling scarce monetary resources toward highly visible projects, overfunding becomes the main failure-cause of underexposed campaigns. This dynamic leads to a *"cannibalization effect"* of overfunded campaigns, that has also been confirmed by Kim, Lee, Cho and Lee (2016), who also state that many promising start-ups in crowdfunding cannot attain their fundraising targets because they are overshadowed by their overly successful counterparts [R.M. Raafat et. al., (2009), A. Moritz et. al., (2015)]. Doshi (2014) even claim that project initiators choose the *"option of not entering"* the platform with an innovative idea, because they fear to be outshined by the *blockbuster* projects.

Moreover, while crowdfunding platforms could generate traffic by promoting overfunded campaigns, the unbalanced allocation of capital could cause *adverse selection* among investors who tend to provide money to already successful projects. This dynamic may deter aspiring entrepreneurs from fundraising because of the fear of being overshadowed, which in turn might reduce these platforms to a *"market for lemons"* (Li Y. et. al., 2020), darkening some high-quality projects. This problem arises since the crowdfunding context imply an *information asymmetry* condition. Typically, information asymmetries emerge when one party involved in an economic transaction possesses greater material knowledge than the other. This usually manifests when the seller of a good or service has greater knowledge than the buyer; however, the reverse dynamic is also possible. In crowdfunding, backers are not perfectly informed about the quality of the different projects, or do not have the capability to distinguish the most innovative and worthy ideas. For this reason, there is the risk on crowdfunding platforms that *lemon* campaigns overshadow *peach* campaigns.

Another interesting point of view in accordance with that of Li Y. and colleagues, is the one of Koch J. and colleagues (2018), who state that massively overfunded projects have been discussed to overshadow other crowdfunding projects which, in turn, receive less money. To solve this problem, in their paper they propose a taxation mechanism to internalize these overfunding externalities, thus improving the overall funding results. So, their evaluation provides evidence that possible modifications of the crowdfunding mechanisms bear the chance to optimize funding results and to alleviate existing flaws.

Also, the contribution of Mollick (2014) confirms the inefficiencies that overfunding can cause on the market, since he found that overfunded projects tend to experience much more delays in delivering the promised rewards.

Another thought is derived from Malave (2012) and his paper; the scholar captures the idea that lowering overfunding could help to fund undervalued projects that otherwise fail to reach their goal. Specifically, it was shown that massively overfunded projects can lead to a positive effect on *"related projects"*. In other words, those projects that have a related topic profit from the existence of overfunded projects (Kim et al., 2016; Liu et al., 2015). However, while *"related projects"* seem to benefit from overfunding, it has been revealed that *"over-successful"* projects "*hurt the performance of less-related projects"* (Liu et al., 2015). As a consequence, a great number of campaigns is adversely influenced concerning the funding performance, since it can be assumed that there are typically more unrelated than related projects on a platform.

2.4.2.2. Overfunding leads to negative implications for the campaign creators

In the previous sub-paragraph, we underlined the fact that overfunding may lead to market inefficiencies, thus causing the market for lemon effect. Another question arises of whether raising extra funds, i.e., overfunding, is beneficial or not for the project creator. The majority of scholars agree that extra funds may create inefficiencies (Makýšová L. et al., 2017; Li Y. et al., 2020; Liu F. et al., Koch J. et al., 2020; Svatopluk Kapounek, Zuzana Kučerová).

Firstly, overfunding can be detrimental for *overly successful* fundraisers because the cash surplus from their overfunded campaigns is unlikely to be effectively managed, as it has not been budgeted in the original business plans (M.A. Stanko et al., 2017). For example, taking into consideration reward-based platforms, a higher level of overfunding is related to the higher costs of the extra rewards that should be delivered to additional project investors: the campaign can cost more because the project creators may not have the capacity to create more rewards and because they may lack the time to deliver them (Makýšová L. et al., 2017). Mollick and colleagues (2014), in accordance with the previous considerations, found that over 75% of reward-based overfunded campaigns significantly fell behind on their production schedule, culminating in protracted delays in product delivery. This could also lead to the violation of promised obligations, damaging entrepreneurs' reputation, and harming their future financing round (M.A. Stanko et al., 2017).

The same happens also when considering the case of an equity crowdfunding platform. Li Y. and colleagues (2020) state that if overfunding emerges, funders' shareholding might be diluted by the additional shares issued by the entrepreneur in exchange for extra monetary resources. This not only shrinks the number of dividends paid to funders, but it could also erode their voting rights given an increased number of shareholders (S.S. Turan, 2015).

In addition, taking into consideration a psychological perspective, overfunding could also pose a threat to *overly successful* fundraisers by inflating their egos and spurring them to take on unnecessary risks. As a consequence, overfunded campaigns tend to become overly ambitious, culminating in unrealistic claims, unfulfilled obligations, and ultimately, disappointed funders (Li Y. et al., 2020).

Lastly, both the papers of Kim et al. (2016) and Li Y. et al. (2020) find indication that some project initiators refuse to start new campaigns on the platform if there are massively overfunded projects.

In contrast to the previously presented points of view, some scholars agree that overfunding can bring positive effects too. Koch J.-A (2018) argues that overfunding can be highly beneficial for those project founders that desire to generate publicity and to sell their products or services. The phenomenon is also relevant for founders that plan a project with so-called stretched goals, e.g., a PC game that will contain more game levels according to the available overfunding. Here, the level of extra funds decides on how much of the additional features can be implemented. Even the platform operators themselves benefit from overfunded campaigns thanks to the additional traffic they bring, the indirect promotion for the platform as well as higher resulting revenues.

2.4.2.3. Overfunding leads to negative implications for backers

As reported in previous sections overfunding leads to market inefficiencies, and in the subsequent part we will dig deeper into the implications that the phenomenon generates on investors who provide their money to support campaigns.

First, it is necessary to mention the study of Du S. and colleagues (2020), who introduce the *overfunding effect*, namely a positive psychological perception that enhances customers' confidence in the product's quality, thus providing positive utilities for the latter. This positive perception in a domain of information asymmetry, where customers do not have the competencies to assess comprehensively and accurately the quality and innovative charge of a project, could lead to a phenomenon of *adverse selection*. Indeed, as also supported by Svatopluk Kapounek and Zuzana Kučerová, overfunding fuels adverse selection in the crowdfunding markets, since it may happen that backers invest their money in low-quality projects only because they are strongly pledged, thus hindering high-quality ones from succeeding.

2.4.3. Factors affecting overfunding

In spite of the previously mentioned inefficiencies associated with overfunding, only a handful of studies have touched upon its main determinants. In the following section we will analyse all the factors studied in the literature as influencing elements of overfunding, emphasizing which correlations (positive or negative influences) have been highlighted by scholars. Following the structure given by Koch A. 2013 in his study, we will divide all the determinants analysed into 5 subcategories: *campaign conditions* (*2.4.3.1*), *project information disclosure* (*2.4.3.2*), *crowd-funder related aspects* (*2.4.3.3*), *platform-related aspects* (*2.4.3.4*), *funding behaviour* (*2.4.3.5*). According to this division, our analysis takes into consideration all the three stakeholder groups contributing to overfunding. Indeed, we will analyse arguments for all of them to have certain egoistic incentives for a further backing of already funded projects instead of prioritizing a more demand-oriented distribution of fundings.

2.4.3.1. Campaign-related conditions

In the following section, we discuss the main campaign-related features associated to overfunding. Specifically, the factors analysed are the *funding target*, the *campaign duration* and the *number and typology of reward levels*.

Regarding the *funding target*, the majority of scholars agree that a higher funding target lowers the probability of reaching overfunding. Precisely, Makýšová L., et al., in 2017 state that an increase in the requested amount is associated with a negative impact on overfunding. In their paper a binary logistic regression was used as a methodology to isolate relevant results explaining project's overfunding. Through this regression method, Makýšová L., et al. found a negative probability correlation between a higher funding target and *"over-success"*.

Similar are the conclusions derived by Koch and colleagues, 2018. Indeed, as main previous researches have shown, the funding goal is an important factor in explaining funding success (Koch and Siering, 2015; Mollick, 2014): the higher the funding goal, the less likely the campaign reaches its target. In his paper, Koch assumes a similar effect on overfunding: the funding goal has a negative influence on project *"over-success"*. This assumption is consistent with the results obtained in the empirical analysis he conducted, based on a linear regression, which showed a negative correlation between overfunding and the funding goal of a campaign.

Consistently, Alessandro Cordova and colleagues, (2015) found that an increase in the project funding goal is connected with a lower probability and extent of success. In particular, by conducting a linear regression on their database, they proved that a 1% increase in the amount requested, reduces the probability the founder will reach or exceed the funding goal.

Differently from the aforementioned scholars, Ma X. and colleagues in their studies of 2018 found that the finishing percentage, namely the absolute funding amount, seems to be higher for projects with higher targets. This result is in contrast with the previous researches, which believe that big-sized projects are faced with more obstacles and are hard to succeed. Actually, at the beginning of their research, Ma X. et al. agreed with the majority of the scholars, defining as hypothesis that the target founding amount has a negative effect on overfunding. But then, after analysing a sample characterised by 64 successful pitches in 2014, 93 pitches in 2015 and 46 pitches in 2016 from January to July on Crowdcube, they stated that higher targets drive the project to be funded more and attract more investors at the later period of crowdfunding. They also agreed that, when the pitch has already reached its goal and accumulated enough money, funders started perceiving a lower risk level associated to their investments, increasing the probability of the pitch to be overfunded.

Another important factor analysed in literature as one of the main determinants of overfunding is the *campaign duration*. In this case, the results obtained by scholars reveal both positive and negative implications on overfunding.

Koch and colleagues, 2018, analysed this factor in their paper, and at the end of the study, they showed that a longer funding period leads to a stronger overfunding probability. In contrast with previous researches (Xiao et al., 2014; Mollick, 2014; Barbi and Bigelli, 2015) they concluded that, in case of successful campaigns, longer funding periods are not interpreted as a signal of bad quality or as a lack of confidence among project founders. Instead, the longer a campaign the higher the opportunity for potential backers to invest in it, thus increasing the probability of reaching "*over-success*".

The same point of view is shared by Alessandro Cordova and colleagues, 2015, who show that project duration increases the chances of success.

On the other way around, Miro Arola in his paper of 2018, contends that the shorter the campaign duration, the stronger is the overfunding probability. The author bases his statement on the behavioural literature highlighting that humans are impatient by nature, characteristic associated with present heuristic for immediate reward (Robson & Samuelsson, 2008). Moreover, he adopted an OLS regression model to evaluate the relationship between overfunding and the different variables, and such analysis confirmed his critical assumption.

The last important factor related to the campaign-related predictors subcategory is represented by the *reward level* offered by the fundraisers. First of all, it is fundamental to specify that in crowdfunding, the difference between the equity model and the reward model revolves around the concept of which is the *"reward"*. In the latter crowdfunding platform, backers and supporters receive a reward based on the amount they have invested in the project, represented (typically) by a product or service. In practice, many times, there is a material presale of a product or a service that will be put into production only later, thanks to the sums collected through the crowdfunding campaign. All in all, in reward-based crowdfunding there is a *"reward"* that backers receive. Instead, in equity crowdfunding platform, according to CONSOB [12] definition, it is possible "*through online investment to buy a real title of participation in a company: in this case, the reward for the financing is the set of property and administrative rights deriving from participation in the enterprise*". For this reason, according to the difference just presented, we divided the thought of the scholars regarding the factor *"reward level"* according to whether it is considered an equity-based platform or a reward-based.

The research of Martínez-Gómez C. et al. of 2020, has provided new insights into the literature of overfunding in equity crowdfunding. Their empirical findings have shown the relevance of campaign features like equity and voting rights to explain overfunding. Their paper embraces previous assertions on the positive signalling roles of low equity and low issuance of shares with voting rights, and they finally state that, first of all, a larger percentage of equity offered will reduce the overfunding of equity crowdfunding campaigns, secondly, offering only shares with voting rights will reduce the *"over-success"* probability. To focus on the overfunding distribution, these scholars applied a quantile regression methodology for a total sample of 299 overfunded campaigns from 2015 to 2018. Overall, empirical results show that the effects of key campaign features (equity, voting rights) are stronger and more significant at the 75th and 90th quantiles for the overfunding level and the number of investors.

Another research which closes the gap in overfunding literature by studying which factors drive the overfunding of projects in equity crowdfunding background is the one provided by Ma X. et al in 2018, since they reflect upon the effect of the *"target equity amount"*. More in depth, these scholars define that target equity offering has a negative effect on project *"over-success"*. In their paper, Ma X. et al refers to a previous study of Mollick (2013), who has pointed out that projects raising a large amount of money are more likely to fail due to the larger funds and investors required, while small-sized

[12] https://www.consob.it/

projects are more likely to succeed. Thus, a similar situation is assumed by the study of Ma X. et al. in overfunding projects.

Taking into account the reward-based crowdfunding platform, the work of Koch in 2018 seems to be the most comprehensive and meaningful. Indeed, in his paper the scholar pointed out that, while funders that focus on rewards tend to increase overfunding, altruistic motives lead to less *"over-success"*. Koch's result can be explained by funders reacting to very inviting rewards. Indeed, the more attractive the rewards are, the more the funders try to profit from them and continue funding in order to get such rewards (even if the campaign has already reached its funding goal). This approach can be defined as a more egoistic behaviour in contrast to the altruistic one of funders donating to campaigns. Koch found the opposite result for projects that invite relatively more altruistic funders: the more a project is funded altruistically, the less it tends to be overfunded. For this reason, Koch concluded that altruistic funders feel good in helping projects to reach their funding goal, but in the moment in which they reach success, altruistic founders feel less emotionally satisfied in funding. So, if a project is already funded, altruistic backers do not longer strongly contribute to further overfunding. In this case, the altruistic founder might search for another project that needs support to reach its goal.

The paper of Miro Arola (2018) takes up the main points of the work of Koch and finally quotes from the author himself *"projects tend to be more overfunded if founders offer more levels of rewards".* Moreover, Miro Arola adds that in the case of equity crowdfunding, stakeholders might behave differently, as the compensation of investment is formulated from equity instead of a product or service reward, thus confirming the distinction previously made at the beginning of the paragraph.

Only the work of Makýšová L. et al. (2017), a comprehensive analysis of reward-based crowdfunding in the Czech Republic, conducted based on the data from 617 projects using the Czech crowd-funding platform Hithit, offers a different point of view, since they stated that certain characteristics of the projects were not proven to be significant across all samples, in particular the number of rewards seems not meaningful to affect overfunding.

Campaign-related factors	***Effect on Overfunding***	***Source***
Funding target	There is a negative correlation between higher funding target and overfunding probability. There is just one study (Ma X. et al., 2018) that in contrast with the extant literature underline the existence of a positive relation between the funding target and the overfunding**.**	• Makýšová L., et al., 2017 • Koch et al., 2018 • Cordova et al., 2015 • Ma X. et al., 2018
Campaign duration	Mixed results.	Longer campaign duration increases the probability of overfunding: • Koch et al., 2018 • Cordova et al., 2015 Shorter campaign duration increases the probability of overfunding: • Miro Arola, 2018
Reward - level	*Equity based platform:* The target equity offering has a negative effect on project overfunding. Moreover, offering only shares with voting rights will reduce the overfunding success of crowdfunding campaigns.	*Equity based platform:* • Martínez-Gómez C. et al., 2020 • Ma X. et al, 2018

	Reward based platform: Mixed results.	*Reward based platform:* While funders that focus on rewards tend to increase overfunding, altruistic motives lead to less overfunding. Projects tend to be more overfunded if founders offer more levels of rewards • Koch, 2018 • Miro Arola, 2018 The number of rewards seems not meaning to affect overfunding. • Makýšová L. et al., 2017

Table 9: Impact of campaign-related factors on overfunding

2.4.3.2. Project Information disclosure

In the paragraph below, we discuss the influence of project information disclosure on overfunding. Indeed, the project funder may disclose different kind of information to potential backers, relying on textual information, like *comments* or *updates* and media-based information like *videos* or *images*.

The fact that the project creator provides more or less information (in the form of *texts*, *pictures*, and *videos*) and communicates or not in an active way, may influence the likelihood to reach overfunding. It has been shown in previous literature that the length of project descriptions (Xiao et al., 2014; Koch and Siering, 2015), the number of pictures (Koch and Siering, 2015), and the provision of video material (Mollick, 2014;

Koch and Siering, 2015; Xiao et al., 2014) have a positive influence on funding success, and it is very interesting to investigate whether the same is found in overfunding.

Starting with the textual information disclosed, Makýšová L. and colleagues, 2017 analysed the effect of *comments* on the overfunding probability. Comments are intended as the number of reviews that the project creator and everyone interested in the idea wrote in the project profile section during the campaign. In the aforementioned paper, it has been proven that a high number of comments reduces the probability of overfunding. In this study, also the effect of the number of updates on overfunding was considered, but they were not proven to affect the phenomenon.

Conversely, Koch and colleagues, 2018 and Miro Arola, 2018 found that textual information has a positive influence on overfunding. Precisely Miro Arola states that a higher number of forum-posts and frequent *pitches updates* published by the project creator could increase the likelihood of overfunding.

Also, media-based information (such as pictures and videos shared) have been studied in literature as factors influencing *"over-success"*.

Ma X. and colleagues, 2018 underline that sharing of *videos* on LinkedIn gets projects exposed to more potential investors and boosts the finishing percentage, thus leading to a higher overfunding probability, while no evidence shows the exposure of videos on Facebook having similar effects. In addition, apart from the number of videos shared, also their length is demonstrated to have a positive impact on overfunding likelihood. Similarly, Cicchiello A.F. and colleagues, 2020 find a positive coefficient but no statistically significant empirical evidence that the presence of a video to support the campaigns positively influences the projects' chance to experience overfunding.

Project information disclosed	***Effect on Overfunding***	***Source***
Textual information	Comments have been proven to have a negative influence on overfunding, while forum post and pitch updates increase the overfunding likelihood. No evidence of the effect of updates on overfunding have been found.	• Makýšová L., et al., 2017 • Koch et al., 2018 • Miro Arola, 2018
Media-based information	There is a positive correlation between the number and the length of videos shared and the overfunding probability.	• Ma X. et al., 2018 • Cicchiello A.F. et al., 2020

Table 10: Impact of project information disclosure on overfunding

2.4.3.3. Fundraisers-related factors

As in the success, also in overfunding case fundraisers' characteristics are fundamental determinants of the phenomenon. Among these factors, the most cited in the literature and, therefore, the ones that we will deeply analyse in the upcoming section are *gender*, *fundraiser's typology* (namely whether the fundraiser is an NGO, a private person or a company), the *fundraiser's social capital*, and finally whether or not the fundraiser can be considered a *"serial"* one.

The *gender* factor has been deeply studied by Cicchiello A.F. and colleagues, 2020. According to their paper the gender of the founders is not significantly related to the likelihood of a campaign to be overfunded, and moreover, the results of the Poisson model they constructed confirm that the gender composition of the team is not significantly related to the phenomenon.

Regarding the *fundraiser's typology*, Makýšová L. and colleagues, 2017 prove that the projects created by the NGOs have lower odds of raising extra funds than projects created by another form of creator, like private people. This finding is consistent with the theory of the two-sided market by Belleflamme et al. (2016), in which potential backers join the market not only for altruistic reasons but also to seek rewards. For this reason, it may happen that a project created by a profit-seeking entity may introduce a more interesting idea and offer more attractive rewards than a project created by an NGO where the rewards or offers are rather symbolic, thus the overfunding odds are lower. In cases in which the fundraiser is a company, different factors come into play. The firms' age and the industry (sector) in which they operate influence the project chance to experience overfunding (Ma X. et al., 2018; Cicchiello A.F. et al., 2020).

In addition, Koch and colleagues 2018 underline that the longer and the more the founder is active on the platform, the stronger is the overfunding. In particular, in the overmentioned paper some considerations are made on the correlation between the founder being a *serial* one, and the overfunding probability, and results show that the more project campaigns a founder has conducted (serial founder), the stronger is the "*over-success*"

Also, the fundraiser's *social capital* influences the overfunding likelihood, and has been analysed by different scholars. Koch and colleagues, 2018; Martínez-Gómez C. and colleagues, 2020, Miro Arola, 2018, converge around the idea that an extensive fundraiser social capital can greatly increase the chances of experiencing overfunding.

Fundraisers – related factors	***Effect on Overfunding***	***Source***
Gender	Gender is not significantly related to the probability of a project of being overfunded.	• Cicchiello A.F. et al., 2020

Fundraisers' typology	NGOs have lower odds of raising extra funds than projects created by another private creator. The firms' age the industry (sector) in which they operate influence the project chance to experience overfunding.	• Ma X. et al., 2018 • Cicchiello A.F. et al., 2020 • Makýšová L., et al., 2017
Serial fundraiser	Projects launched by a serial fundraiser experience higher probability of being overfunded.	• Koch A. et al., 2018
Social capital	The fundraisers' social capital positively influences the overfunding likelihood.	• Koch A. et al., 2018 • Martínez-Gómez C. et al., 2020

Table 11: Impact of fundraisers-related factors on overfunding

2.4.3.4. Platform-related factors

The literature concerning the determinants affecting overfunding and related to the characteristics of the platform, is not widely extensive. Indeed, the only factors considered are the *platform typology* (whether equity or reward based) and the *projects quality* driven by the platform itself.

Referring to the paper of Miro Arola (2018), based on 185 campaigns from the opening of the platform *"Invesdor Oy"* (a dominant equity crowdfunding platform in the Nordic countries) in 2012 up until September 2017, overfunding seems to be a more pronounced phenomenon in equity rather than reward-based crowdfunding. Indeed, as also mentioned in the previous paragraph, the work of Arola resumes the point of view of Koch (2016 and 2018), but underlines that the different structure of the platform (equity or reward-based) could be significant to affect overfunding.

Another interesting point is faced by Koch in his paper. Indeed, the scholar states that the *typology of the platform* can influence funding results. Furthermore, the author contends that if a project campaign promises to contribute substantially to the platform's revenue, the platform itself could place this project more prominently on the websites or indicate quality by labels to further support its overfunding (Koch 2016 and 2018). Indeed, platform operators especially benefit from projects that are characterized by a relatively fast-growing sum of pledged money (Agrawal et al., 2013). Thus, operators principally have an incentive to support fast-growing projects, that, for this reason, are better positioned on the website. Such a prominent placement promotes further fundings (Do et al., 2012) and increases platform's revenues. In conclusion, Koch pointed out that the platform is able to drive overfunding by providing a better positioning on the website to high-quality projects.

This idea is also coherent with the study of Ma X. et al. (2018) who also underline that *project quality* will not only attract more supporters but also encourage existing ones to promote such project to other potential investors or external media, thus, increasing the popularity of the campaign and the extent of overfunding. For this reason, if the platform reveals and allows funders to access project's quality, the overfunding effect increases.

Platform– related factors	***Effect on Overfunding***	***Source***
Platform typology (equity or reward based)	Overfunding seems to be a more pronounced phenomenon in equity- than reward-based crowdfunding	• Miro Arola, 2018

Projects quality driven by platform	The platform could increase the popularity of the project and the extent of overfunding by revealing to fundraisers the quality of the project itself.	• Koch, 2018 • Ma X. et al., 2018

Table 12: Impact of platform-related factors on overfunding

2.4.3.5. Funding behaviour

As reported in paragraph *2.3.1.3*, backers play a crucial role in determining crowdfunding success, and it is very interesting to understand whether the same characteristics and funding dynamics influence also the overfunding phenomenon. The main factors studied are the following: *timing and funding dynamics, number of backers* and mean donation per backers and the so-called *herding behaviour.*

In their study, Yen C.-H., Lee Y.-C., Fu W.-T (2020) point out that on existing crowdfunding platforms, the allocation of money is often not regulated, which leads to a less-than-ideal distribution of resources. More in depth, these scholars take into consideration different *dynamics and timing*: first, the rate of new donations tends to increase when project deadlines are close (Agrawal, 2013), and when a project is near to its target, suggesting that, all else being equal, donors prefer projects that are more likely to succeed. This result is supported also by the study of Ma. X. (2018), saying that the fact that the project has already been successful serves as a more certain and stronger signal of project quality, so convincing more backers to participate and contribute. So, the results of Yen C.-H. suggest that donation dynamics and timings may play a significant role in influencing the allocation of donations, even if it leads to higher inequality and unpredictability of outcomes, as in the case of overfunding.

Makýšová L. and colleagues, 2017; Miro Arola, 2018, Cordova and colleagues, 2015, found that the *number of backers* and the *mean contribution per backer* have a positive impact on the overfunding probability. Precisely in the case of Makýšová L. and colleagues, 2017 when only an odd ratio for the mean contribution per funder is considered, it implies a negative effect. However, the variable is also included in the interaction with the requested amount; therefore, both coefficients are counted together, showing the positive effect on campaign overfunding.

Lastly, the impact of the *herding behaviour* has been analysed. Like Raafat et al., 2009, we refer to herding *as the alignment of the thoughts or behaviours of individuals in a group through local interaction and without centralized coordination.*

In particular, the study of Li Y. and colleagues, 2020, focuses on the impact of initial herd on overfunding, under three dimensions: *maturity, intensity and persistency.* The results show that:

- The *maturity* of an initial herd (i.e., an indication of the time it takes for the initial herd to form) in a campaign is negatively associated with its degree of overfunding.
- The *intensity* of an initial herd (i.e., the number of funders it galvanizes at its peak) in a campaign is positively associated with its degree of overfunding.
- The *persistency* of an initial herd (i.e., how long the initial herd can persist) in a campaign is positively associated with its degree of overfunding

In addition, and consistently with the aforementioned study, Miro Arola, 2018 and Svatopluk Kapounek, Zuzana Kučerová, 2019, argue that the herding behaviour will positively affect overfunding.

Funding behaviour	***Effect on Overfunding***	***Source***
Dynamics and timing	Donation dynamics and timings may play a significant role in influencing the allocation of fundings as in the case of overfunding phenomenon.	• Yen C.-H. et al., 2020 • Ma X. et al., 2020
Number of backers and mean contribution per backer	Scholars show positive correlations among these factors and the overfunding probability.	• Miro Arola, 2018 • Cordova et al., 2015 • Makýšová L., et al., 2017
Herding behaviour	Herding and initial behaviour will positively affect overfunding.	• Li Y. et al., 2020 • Miro Arola, 2018 • Svatopluk Kapounek, Zuzana Kučerová, 2019

Table 13: Impact of funding behaviour on overfunding

Chapter 3 - Objectives of the thesis

3.1. Research gaps

As reported in previous chapters, the literature on crowdfunding has been widely studied in several respects. In particular, the factors determining the success of a campaign have been investigated, considering both static and dynamic variables and emphasizing the importance of considering these success factors during the duration of the campaign.

Instead, the theme of overfunding, and therefore the ability of a campaign to get more than required (*total pledge* > *target*), still remains on hold in existing literature.

In particular scholars have focused their attention only on three main topics, as reported in the literature review:

1. Overfunding as a phenomenon peculiar of crowdfunding which is not found in other financing methods (Business Angels, Venture Capitalists);
2. Overfunding as a source of market inefficiencies, analysing the *market for lemon* issue and the negative implications overfunding can cause to all the crowdfunding stakeholders;
3. Overfunding determinants, focusing on which are the factors leading to this phenomenon.

As already pointed out, the literature on overfunding, apart from the aforementioned topics, is still young and with many gaps to be covered, which we believe, however, fundamental for an overall understanding of the phenomenon. For this reason, the proposed thesis aims to deepen this theme, trying to investigate the main unresolved areas regarding *"over-success"*.

The following section is organised as follows: first of all, we point out that the literature lacks an integrated definition of overfunding, without clearly discriminating it from the *"simple"* success.

Moreover, as said, the extant literature focuses on a single or limited number of factors influencing the phenomenon, without providing an integrated overview of all the determinants and differentiating them by level of importance. Indeed, all the scholars exhibit the factors considering them all having the same impact on overfunding and without discriminating between their influences on the phenomenon. Therefore, in our thesis we focus on the need of providing a comprehensive overview of all the factors causing overfunding and assessing the relative importance of both static and dynamic factors to predict it.

Subsequently, we address the impact of the time variable, namely which time horizon should be considered as crucial and insightful to determine and predict overfunding.

Finally, we define our research questions and summarize the objectives of our work.

3.1.1. Analysis of overfunding definition and operationalisation

First of all, it is very important to underline the difference between the concept of *definition* and *operationalization*, since in most of the cases they are considered as overlapping. Definition is the process of conceptualising or specifying concepts, while operationalisation is the process by which a researcher specifies precisely how a concept will be measured (Vito La Vecchia, Politecnico di Bari[13]).

In the following section, we will firstly address all the definitions provided in literature, then we will focus on the operationalisations, underlying their limitations.

[13] Differenza tra concettualizzazione e operazionalizzazione | Informatica e Ingegneria Online (vitolavecchia.altervista.org)

The overfunding phenomenon happens because, once a specific project gains full funding, the financing campaign is not closed, thus the raised sum of money can exceed the pre-set goal, and other lenders may still enter this auction and contribute in *"overly successful campaigns"*; as such, the project can be overfunded (Koch, 2016; Li Y et al., 2020).

Following these dynamics, in literature the term overfunding has been defined in slightly different ways by scholars. Deepening, Ma. X. et al., 2018, conceptualised the phenomenon stating that it is: *"the fact that the project has already reached its funding goal"*. This statement is very similar to the one provided by Martínez-Gómez C et al., 2020, who stated that overfunding happens when *"the projects successfully surpass the fundraising goal stipulated by fundraisers"*.

Moreover, according to Koch and colleagues, 2018, the term overfunding has been used by scholars with two different shadows, precisely in two similar, but not equal, context, that in practice need to be differentiated. Firstly, a project is called overfunded the moment its funding exceeds the goal (e.g., Frydrych et al., 2014; Ma X. et al., 2018; Li Y. et al., 2020; Cordova et al., 2015). Secondly, the term can be especially used when a project's funding is considerably higher than its funding goal (e.g., Mollick, 2014).

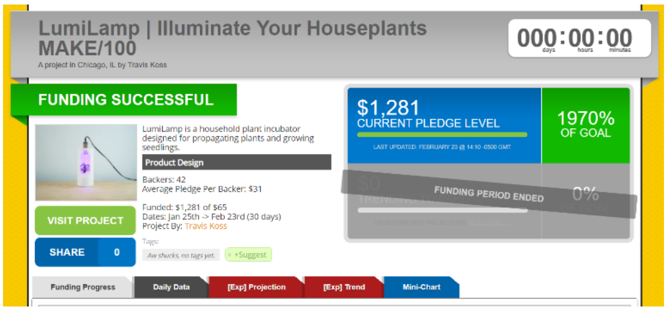
In these lines of reasoning, we find a number of limitations. In the first instance, by adopting a definition such as the ones mentioned above, the boundaries between the concept of success and overfunding become unclear, making it difficult to understand the phenomenon and its consequences.

Indeed, it is very rare for a successful campaign to reach a pledge exactly equal to the target (*pledge = target*). In the majority of the cases, a successful campaign exceeds the initial funding goal, because it may happen that the final pledge provided by the last backer brings the campaign to pass the target (even of one dollar). In these cases, which represents the majority of the instances for crowdsourcing, it is difficult to understand whether the campaign is simply a successful one or is *"over-successful"*.

This is why it is necessary to define a threshold, that is, what percentage the pledge must exceed the target in order to discriminate between success and overfunding.

To better explain the concept just mentioned, below, we report an example with two different campaigns published on the reward-based platform Kickstarter, almost characterised by the same duration (30 and 34 days). Both projects have had a positive outcome as they both reached the target amount set. More precisely, both campaigns, *Lumilamp* and *Grano*, have achieved a pledge greater than the goal required to launch the two projects: the first has collected about more than the double of the goal, the second instead only 7 dollars more. So, it is interesting to dig deeper these two campaigns in order to understand which one is a *"successful"* campaign, and which one is an *"overfunded"* one, since according to the definitions available in the literature, this is not absolutely clear.

Examples:



LumiLamp | Illuminate Your Houseplants MAKE/100

A project in Chicago, IL by Travis Koss

FUNDING SUCCESSFUL

LumiLamp is a household plant incubator designed for propagating plants and growing seedlings.

Product Design

Backers: 42
Average Pledge Per Backer: $31

Funded: $1,281 of $65
Dates: Jan 25th -> Feb 23rd (30 days)
Project By: Travis Koss

Tags:
Aw shucks, no tags yet.

$1,281 CURRENT PLEDGE LEVEL

LAST UPDATED: FEBRUARY 23 @ 14:10 -0500 GMT

1970% OF GOAL

FUNDING PERIOD ENDED

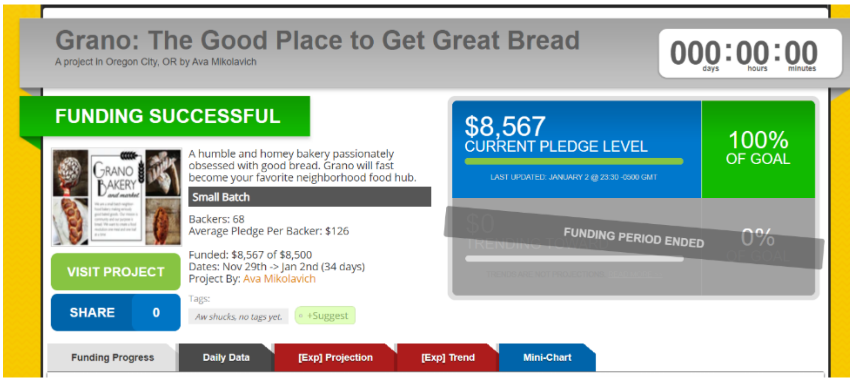
Figure 5: LumiLamp project

LumiLamp [14]is a minimal 3D printed grow pod designed for homes. It is shaped for growing seeds and plant trimmings indoors in any season. As reported in the main page of Kickstarter, the campaign created by Travis Koss:

[14] https://www.kicktraq.com/projects/1138486685/lumilamp-illuminate-your-houseplants-make-100/

- Has a duration of 30 days (from Jan 25th to Feb 23rd)
- Has received an average pledge per backer of $31
- Has been financed by 42 backers

In its lifetime-period, the campaign received more than $1.280, even if the goal set by the creator Koss was $565, so more than the double of what was required.



Grano: The Good Place to Get Great Bread
A project In Oregon City, OR by Ava Mikolavich

000:00:00
days hours minutes

FUNDING SUCCESSFUL

A humble and homey bakery passionately obsessed with good bread. Grano will fast become your favorite neighborhood food hub.

Small Batch

Backers: 68
Average Pledge Per Backer: $126

Funded: $8,567 of $8,500
Dates: Nov 29th -> Jan 2nd (34 days)
Project By: Ava Mikolavich

Tags:
Aw shucks, no tags yet. +Suggest

$8,567
CURRENT PLEDGE LEVEL
LAST UPDATED: JANUARY 2 @ 23:30 -0500 GMT

100%
OF GOAL

FUNDING PERIOD ENDED

Figure 6: Grano project

As reported in the description of *"Grano"* [15]: *"The good place to get great bread"*, the project came alive thanks to a small batch neighbourhood bakery, making seriously good baked goods. The mission of the project is to create a community, promoting a food revolution one meal and one loaf at a time. The campaign created by Ava Mikolavich:

- Has a duration of 34 days (from Nov 29th to Jan 2nd)
- Has received an average pledge per backer of $126
- Has been financed by 68 backers

In its lifetime-period, the campaign received more than $8.567, basically $7 dollars more than required, since the goal set by the creator was $8.560.

[15] https://www.kicktraq.com/projects/1061508054/grano-the-good-place-to-get-great-bread/

Therefore, as it emerges from these two campaigns, it is not clear which of them can be considered successful or *over-successful* or even if both can be considered overfunded, because, as the literature claims, the two projects exceed the funding target. For this reason, we believe that a simple qualitative consideration, namely the fact that the pledge must exceed the target, is not enough to define the phenomenon and distinguish it from success. Thus, it is necessary to make quantitative considerations and introduce a concrete and consistent operationalization of the overfunding.

In literature, indeed, just few scholars provided a way to measure the overfunding variable.

In this context, a very relevant study is the one from Koch and colleagues, 2018, who examined the distribution of the amount of funding that crowdfunding campaigns have achieved, setting the finally reached funding X into relation to the initially defined funding goal X_{goal} of the project as $R_{fund} = X / X_{goal}$. For unsuccessfully funded projects, this funding extent R_{fund} ranges always in the interval of [0, 1[. Successfully funded projects always have a $R_{fund} \geq 1$, while a value of $R_{fund} = 1$ means that a project has exactly reached the funding goal. They defined a project to be overfunded if $R_{fund} > 1$.

Other scholars (Makýšová L., 2018; Vaceková G., 2018; Ma. X. et al., 2018; Li Y. et al., 2020) operationalised the variable as the ratio between absolute funding amount and target funding amount between 1 and higher.

We contend that it is important to identify a more precise threshold, because, stating that a project is overfunded when the ratio between pledge and target is higher than 1 is still too general, and gives no insights of how to distinguish "*over-success*" and success. It is necessary to identify of which percentage the aforementioned ratio must exceed 1 to be able to consider a project as overfunded.

Definition	***Source***
The project successfully reaches and exceeds the funding target	• Koch, 2016 • Li Y et al., 2020 • Frydrych et al., 2014 • Ma X. et al., 2018 • Cordova et al., 2015
A project's funding is considerably higher than its funding goal	• Mollick, 2014

Table 14: Overfunding definition

Operationalization	***Source***
Ratio between absolute funding amount and target funding amount between 1 and higher.	• Makýšová L. et al., 2018 • Vaceková G. et al., 2018 • Ma. X. et al., 2018 • Li Y. et al., 2020
A project is said to be overfunded if $R_{fund} X / X_{goal} > 1$	• Koch et al., 2018

Table 15: Overfunding operationalization

3.1.2. Factors influencing overfunding

As mentioned at the beginning of the paragraph, scholars have focused on studying the factors predicting the success of a campaign, and the existing works, despite trying to analyse the determinants of overfunding, have always focused on a single or limited number of factors, without assessing the relative importance and influence of each of them on the phenomenon under analysis. We believe that a broader spectrum of factors can provide a more integrated view of the impact that the choices undergone by fundraisers have on the overfunding of their campaigns. Moreover, the extant studies, portraying a survey of the determinants of overfunding, do not disclose any information regarding the relative importance of these factors in predicting whether the campaign will be successful or overfunded. The possibility to investigate the relative importance of overfunding factors (both static and dynamic ones) together with their actual effect on the campaign, could allow for a further development of the current literature, but also provide meaningful insights for practical purposes.

3.1.3. Impact of the time horizon

The literature on success in its several studies has analysed how the time affects crowdfunding success, discovering that the first week of the campaigns is essential for predicting the funding outcome. Colombo and colleagues (2015) find that early contributions are fundamental for determining campaign's success using a dataset collected from Kickstarter. Indeed, the first days of a campaign are crucial given that the level of uncertainty is high, and it is essential to trigger a *"success-breeds-success"* process (Colombo et al., 2015b).

The literature on overfunding has not yet focused on understanding whether, as in the case of success, it is sufficient to consider only the first week as the most critical period, or it is needed to consider a broader time horizon to have a complete overview of the phenomenon.

3.2. The objectives of the thesis

To sum up, the research objectives of our thesis start from the desire to integrate the extant knowledge on overfunding, and trying to clearly differentiate it from success.

To this end, we aim at providing a definition and operationalization of the phenomenon, taking a step forward from existing literature, which remains very vague on the subject, defining it as *"pledge higher than target"* (Ma et al., 2017; Li et al., 2020).

O1 *– Provide a definition and operationalization of the overfunding concept and define a threshold that discriminates between success and "over-success".*

In addition, our analysis aims to investigate whether, as in the case of success, it is possible to predict overfunding based on static and dynamic characteristics, and which is the relative importance of the above-mentioned factors in causing the phenomenon.

O2 *– Assess which static and dynamic variables can predict and influence overfunding and which are the most relevant ones.*

As a last point, the time horizon will be taken into account, investigating which period during the campaign duration is the most critical for the overfunding phenomenon. In particular, as it has already been analysed in the previous literature, the first week seems to be the decisive one to decide the success or failure of a campaign. Our goal is to understand if the same result could be applied to overfunding and so, if the first week is crucial to obtain more than simply success, or if a broader time spectrum should be considered.

O3 *– Identify which period of the campaign appears to be the most critical to predict and achieve overfunding.*

To conduct this analysis, we used a sample including 352 campaigns launched on Kickstarter between 2016 and 2017 and we develop a set of classifiers derived from machine learning techniques. Our classifiers consist of 20 static predictors and 4 dynamic ones, chosen on the basis of the factors that crowdfunding literature has shown as influencing the outcome of a campaign.

Chapter 4 – Machine learning theory

After defining the main research objectives that our thesis proposes to achieve on the theme of overfunding, it is now necessary to specify the methodology followed to provide a meaningful response to all the objectives set. In particular, the approach is based on a *machine learning technique*, specifically a *supervised learning* algorithm that uses a model to learn a mapping between input examples and the target variable. According to this methodology, the next paragraph will be fully dedicated to give a general overview of the main features of machine learning, its main applications and all the main algorithms with all their different combining techniques. This in-depth analysis will clarify the whole process that will be addressed in the next sections of the thesis.

4.1. Machine learning

As said, before introducing all the possible classification methods and combinations that the discipline based on AI provides, it is worthy to underline what is meant by the term *"machine learning"*, giving a definition of the subject, a general overview with its fields of application and emphasizing its main learning problems.

Defining the terms *"machine"* and *"learning"* in the ICT (Information and Communication Technology) context, will be a guideline to fully understand the discipline and its main functions. As reported by Omary Z. et al. (2010) according to Oxford English Dictionary, "*the computer is a machine for performing or facilitating calculation; it accepts data, manipulates them and produces output information based on a sequence of instructions on how the data has to be processed*". Additionally, learning can be defined as a process of acquiring modifications in existing skills, knowledge and habits through experience, exercise and practice. Moreover, for the last definition, Witten and Frank (2005) state that *"things learn when they change their behaviour in a way that makes them perform better in the future"*. Therefore, a general but complete definition of machine learning should involve the knowledge acquisition process and define where the type of knowledge could be obtained.

4.1.1. A general overview

Machine learning is a subfield of Artificial Intelligence, and its first definition was provided in the 1950s by an AI pioneer Arthur Samuel as *"the field of study that gives computers the ability to learn without explicitly being programmed."* It is focused on teaching computers to learn from data and to improve with experience, and this is the reason why it is different from traditional programming. In machine learning, algorithms are trained to find patterns and correlations in large data sets and to make the best decisions and predictions based on that analysis. Machine learning applications improve with use and become more accurate the more data they have access to.

The basic process is to give *training data* to a *learning algorithm* which, then, generates a new set of rules, based on inferences from the data: this is in essence generating a new algorithm, formally referred to as the machine learning model. By using different training data, the same learning algorithm could be used to generate different models.

Inferring new instructions from data is the core strength of machine learning. It also highlights the critical role of data: the more data available to train the algorithm, the more it learns. In fact, many recent advances in AI have been made possible by the enormous amount of data enabled by the *Internet of Things*.

4.1.2. Machine Learning applications

Machine learning, as we are going to explain in this paragraph, has several practical applications, in many different fields. Indeed, thanks to its capability of making predictions, machine learning analysis and classification is applied by companies who need to deal with huge amount of data to run their businesses and get an edge over their competitors. These companies belong to a variety of sectors, such as robotics, finance, marketing, healthcare and education, among others. In this paragraph we provide a brief overview of the most recent and most common applications in some of these industries.

The *healthcare sector* has always been an early adopter of technological advances. These days, machine learning plays a key role in many health-related realms, including the development of new medical procedures, the handling of patient data and records and the treatment of chronic diseases. For example, there are machine learning techniques such as *K-Nearest Neighbour, Logistic Regression, Decision Tree, Support Vector Machine,* and *Deep Learning* (Supriya and Deepa, 2020), able to identify tumours and customize clinical treatment or also magnetic resonance imaging analysis to identify Alzheimer's 15 years before the first symptoms arise.

Another interesting application sector of machine learning algorithms, is the *educational* one, where *Bayes Classifiers, K-Nearest Neighbour* and *Decisions Trees* can be used to estimate the final grade of students (Minaei-Bidgoli et al., 2003), or to assess student's satisfaction (Thomas and Galambos, 2004).

A further application sector, is the *digital marketing* one, characterized by a recent use of the machine learning techniques. Some of the applications adopted are: *chatbots* used to improve customer service; personalized product recommendation, based on your previous searches; dynamic pricing and sentiment analysis. The latter is adopted in a variety of fields, including financial and retail ones, and is based on *Natural Language Processing* (*NLP*) techniques, in order to capture and interpret customers' emotions in front of a product, brand or advertisement.

The last sector worth mentioning is the *automotive* one, which is making machine learning and artificial intellect the banner of a new era of transportation. Thanks to computer vision techniques cars are now able to autonomously detect pedestrians, horizontal and vertical signs, distance, even in less-than-optimal conditions.

4.1.3. Machine learning categories

Machine learning is comprised of different types of models, using various algorithmic techniques. Depending upon the nature of the data and the desired outcome, one of three learning models can be used: *supervised, unsupervised, or reinforcement learning*. These three learning categories are associated, as said, with different machine learning algorithms that represent how the learning method works.

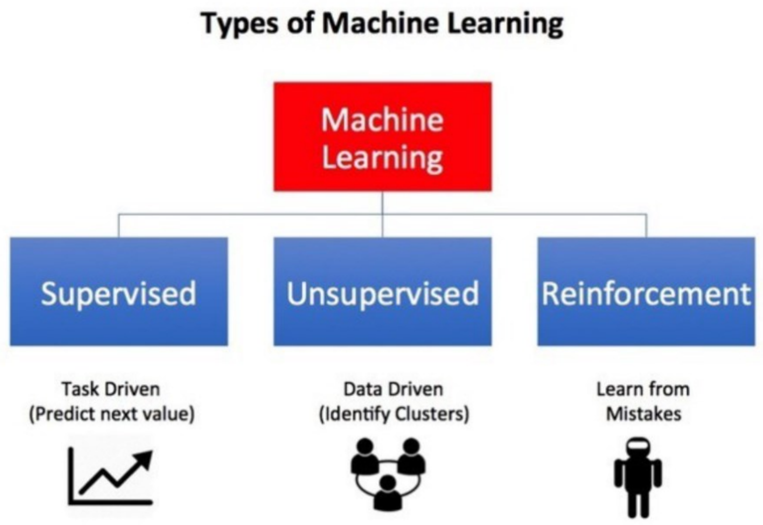

Types of Machine Learning
Machine Learning
Supervised
Unsupervised
Reinforcement
Task Driven (Predict next value)
Data Driven (Identify Clusters)
Learn from Mistakes

Figure 7: Machine learning types. Source: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f

Christopher M. Bishop stated in his book *"Pattern recognition and machine learning"* (2006) that "*Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems"*. In supervised learning algorithms, the machine is taught by example. Supervised learning models consist of *"input"* and *"output"* data pairs, where the output is labelled with the desired value. Indeed, this typology of algorithm analyses the training data and processes an inferred function, which is used as a starting point for mapping new samples. An optimal scenario will allow the function produced by the algorithm to generalize from the training data and effectively determine the class labels for unseen instances and situations. Basically, there are two main types of supervised learning problems: they are *classification* that involves predicting a class label and *regression* that implies a prediction of a numerical value. For these typologies of problems, different kind of *supervised* algorithms exist, popular examples include: *Decision Trees, Support Vector Machines*, and many more.

In unsupervised learning models, there is no answer key. The machine studies the input data (much of which is unlabelled and unstructured) and begins to identify patterns and correlations, using all the relevant, accessible data. Basically, *"unsupervised learning"* one "*there is no instructor or teacher, and the algorithm must learn to make sense of the data without this guide."* (Deep Learning, pg. 105, 2016). Indeed, it operates only upon the input data without target variables or outcomes. As such, unsupervised learning does not have a *"teacher"* correcting the model, as in the case of supervised learning. In many ways, unsupervised process is modelled on how humans observe the world. Human beings use intuition and experience to group things together, and basically, as they experience more and more examples of something, their ability to categorize and identify it becomes increasingly accurate. There are two main problems faced through unsupervised machine learning: the first one is *clustering*, which is based on finding groups in the data. Deepening, cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering. Instead, the second unsupervised problem is known as *density estimation* and involves summarizing the distribution of data.

In reinforcement learning, the algorithm interacts with a dynamic environment, which provides to the agent feedbacks in terms of rewards or punishment. Specifically, the use of an environment means that there is no fixed training dataset, rather a single goal or set of goals that an agent is required to reach and feedbacks about performance toward the goal. Indeed, the reinforcement learning model does not include an answer key but, rather, inputs a set of allowable actions, rules, and potential end states. When the desired goal of the algorithm is fixed or binary, machines can learn by example. But in cases where the desired outcome is mutable, the system must learn by experience and reward. In reinforcement learning models, the *"reward"* is numerical and is programmed into the algorithm as something the system seeks to collect. Some popular examples of reinforcement learning algorithms include *Q-learning, temporal-difference learning, and deep reinforcement learning.*

4.2. Supervised machine learning algorithms

Machine learning algorithms allow computers to train on data inputs and use statistical analysis to predict output values within an acceptable range. For this reason, machine learning helps computers to build models from data examples, so that they can create and structure an automatic decision system based on the received inputs. As new data is fed to these algorithms, they learn and optimize their operations to improve performance, developing intelligence over time.

This section will be divided into two subparagraphs. The first one will be fully dedicated to provide a synthetized overview about the machine learning supervised techniques. To this end, we follow the classification illustrated by Kotsiantis and colleagues (2006), who described various algorithms and the recent attempt for improving classification accuracy, namely ensembles of classifiers. According to the various typology of supervised algorithm studied by these scholars, we will discuss in the following paragraph *perceptron-based algorithms*, then, *statistical learning algorithms, support vector machines and logic-based algorithms*. For the latter, we will dig deeper on the topic of *decision trees*, since they will be fundamental to further develop and conduct our analysis. In the second part, we will describe and define all the steps to be followed when performing a classification analysis. In particular, we will discuss about the training and testing phases and we will define the most common parameters to assess the performance of a classification model.

4.2.1. The perceptron-based techniques

In machine learning, the *perceptron* is an algorithm for supervised learning of binary classifiers, developed by Rosenblatt between 1950 and 1960 (Rosenblatt, 1958). According to this model, a single neuron detects whether any function is an input or not and classifies them in either of the classes. Deepening, a perceptron, which represents a *biological-neuron* of the human brain, acts as an *artificial-neuron* that performs human-like brain functions: as said, it conducts two-class classification and enables neurons to apprehend and register information procured from the inputs.

Given the input feature values, x_1 through x_n and given the connection weights/prediction vector (typically real numbers in the interval [−1, 1]), w_1 through w_n, the perceptron computes the sum of weighted inputs: $\Sigma_i x_i w_i$ and its output is binary (i.e., the class label of the observation):

$$f(x) = \begin{cases} 1 & if \sum_{i}^{n} x_i w_i > threshold \\ 0 & otherwise \end{cases}$$

As shown in the formula, the output goes through an adjustable threshold, and its value is fixed in order to consider bias. The weights w_i connected to the features' value are updated while the algorithm is trained based on the class label of the observations provided. Since the class is the output value of the model, each time we observe a positive class (i.e., the class associated with the value of 1) the parameters are updated.

The training algorithm of the perceptron works online: the model is updated by taking into account a single data instance. Applying this update for all the data in the training set results in a so-called *epoch* of the algorithm. Training a perceptron typically requires multiple epochs. The prediction rule obtained through the training is then used for predicting the class label on the test set.

Moreover, the perceptron learning algorithm does not terminate if the learning set of data is not linearly separable; indeed, if the vectors are not linearly separable, learning will never reach a point where all vectors are classified properly (*single-layer perceptron model*).

When the data are not linearly separable, the multi-layer perceptron, also known as *Artificial Neural Networks* (ANNs) can be used. These multi-layer neural networks consist of a computational model that reproduces the dynamics of natural neurons: such neurons receive signals through synapses, when the signal is above a certain threshold, the neuron is activated and it emits a signal through the axon (Gershenson, 2003). This multi-layer network consists of a large number of units (neurons) joined together in a pattern of connections (Figure 8). Units in such networks are usually divided into three classes: *input units*, which receive information to be processed; *output units*, where the results of the processing are found; and units in between known as *hidden units*. These three classes correspond to the layers that constitute the network, i.e., *input, output and hidden layers.*

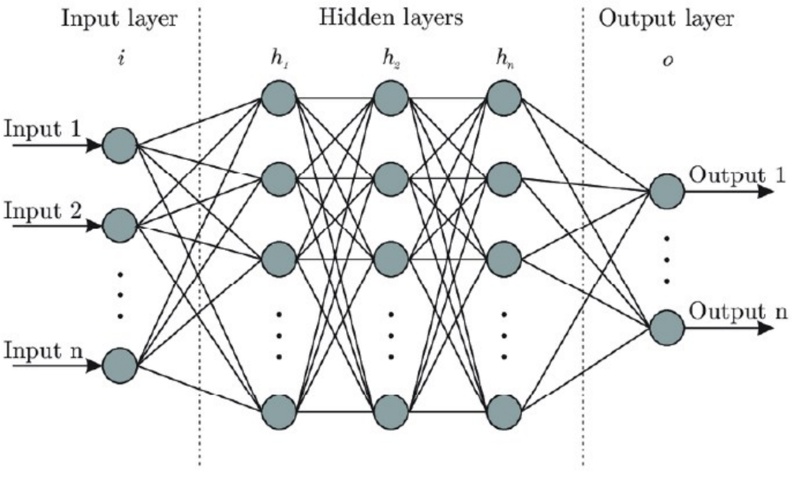


Figure 8: Machine learning layers

Feed-forward ANNs allow signals to travel one way only, from input to output. Such Networks can be viewed as weighted directed graphs, where the nodes are formed by the artificial neurons and the connection between the neuron outputs and neuron inputs can be represented by the directed edges with weights.

Firstly, the *Artificial Neural Network* receives the input signal from the external world in the form of a pattern and image in the form of a vector. These inputs are then mathematically designated by the notations $x_{(n)}$ for every n number of inputs. Then, each of the input is multiplied by its corresponding weights (these weights are the details used by the artificial neural networks to solve a certain problem). In general terms, these weights typically represent the strength of the interconnection among neurons inside the artificial neural network. All the weighted inputs are summed up inside the computing unit (yet another artificial neuron). During classification, the signal at the input units propagates all the way through the net to determine the activation values at all the output units. Each input unit has an activation value that represents some features external to the net. Then, every input unit sends its activation value to each of the hidden units to which it is connected. Each of these hidden units calculates its own activation value and this signal are then passed on to output units. The activation value for each receiving unit is calculated according to a simple activation function. Such function sums together the contributions of all sending units, where the contribution of a unit is defined as the weight of the connection between the sending and receiving units multiplied by the sending unit's activation value (Kotsiantis et al., 2006).

4.2.2. Statistical learning algorithms

Differently from Artificial Neural Networks, which are perception-based models that provide a classification, statistical approaches are characterized by having an explicit underlying probability model, which basically provides also a probability that an instance belongs in each class. Indeed, according to Kostiantis and colleagues (2006), the main scholars of reference, this explicit underling probability model identifies the characteristics of the data in the sample and links them to different classes, so that it is possible to detect for new observations their probabilities of belonging to each of the previously considered classes.

In this category of algorithms, *Bayesian networks* and *instance-based methods* are included.

Bayesian Networks

Over the last 20 years or so, *Bayesian networks* (BNs) have become the key method for representation and reasoning under uncertainty in AI. BNs not only provide a natural and compact way to encode exponentially sized joint probability distributions, but also provide a basis for efficient probabilistic inference (Guo and Hsu, 2002).

A Bayesian Network (BN) is a widely-used class of graphical model for probability relationships among a set of variables (features). This model consists of two parts: a structure and parameters. The structure *S* is a directed acyclic graph (*DAG*) and the nodes in *S* are in one-to-one correspondence with the features *X*. The arcs represent casual dependencies among the features while the lack of possible arcs in *S* has the meaning of conditional independencies; the nodes represent the features. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents. The parameters, instead, consist of conditional probability distributions associated with each node. Thus, figure 9 shows a fully specified Bayesian network:

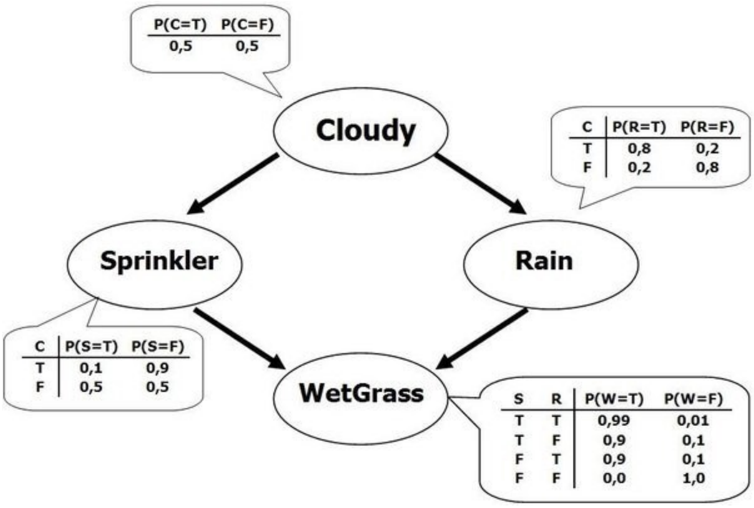


Figure 9: S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall Series in Artificial Intelligence, 1995.

To specify the probability distribution of a Bayesian network, one must give the prior probabilities of all root nodes (nodes with no predecessors) and the conditional probabilities of all non-root nodes given all possible combinations of their direct predecessors (Charniak, 1991). In the end, Bayesian networks allow one to calculate the conditional probabilities of the nodes in the network, given that the values of some of the nodes have been observed. To take the earlier example proposed for the first time in 1995 by S. Russell and Norvig, considering jointly the probability of raining and the one of activating the sprinkler, it is possible to calculate the conditional probability of wet grass given these pieces of evidence.

Within the general framework of inducing Bayesian networks, there are two scenarios: known structure and unknown structure. In the first one, the structure of the network is given and assumed to be correct. Once the network structure is fixed, each node in the network has an associated *Conditional Probability Table* that describes the conditional probability distribution of that node given the different values of its parents. If the structure is unknown, one approach is to introduce a scoring function (or a score) that evaluates the *"fitness"* of networks with respect to the training data, and then to search for the best network according to this score (Kotsiantis et al., 2006).

In spite of the remarkable power of Bayesian Networks, they have an inherent limitation. Indeed, they are not suitable for datasets with many features (Cheng et al., 2002). The reason for this is that trying to construct a very large network is simply not feasible in terms of time and space. This is the reason why in many instances *Naïve Bayesian Networks* are preferred to Bayesian Networks. The formers are very simple Bayesian networks which are composed of acyclic graphs (DAGs) with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes. This assumption is wrong in the majority of the cases and this is why Naive Bayesian classifiers are usually considered less accurate that other more sophisticated learning algorithms (Kotsiantis et al., 2006).

Instance – Based Learning

In machine learning, *instance-based learning*, also known as *memory-based learning,* is a family of algorithms that works comparing completely new instances of problems with events and situations seen in training, which have been registered in memory. Since computation is delayed until a new instance is observed and analysed, these algorithms are sometimes referred to as *"lazy"*. Going deep on the functionalities of these algorithms, each instance is a set of attribute-value pairs and all samples are assumed to be described by the same set of n attributes (Aha et al., 1991). In particular, only one attribute indicates the class of reference, and so it is considered the *"category attribute"*, while all the others that characterize a particular instance or event are utilised as the *"predictor attributes"* to perform the classification by the algorithm.

One of the most common application of the Instance-Based Learning algorithms is represented by the *k-Nearest Neighbour* (kNN) algorithm, which is based on the principle that the instances within a dataset will generally exist in close proximity to other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbours. The kNN locates the *k* nearest instances to the query ones and determines their class by identifying the single most frequent label (Kotsiantis et al., 2006). It is important to underline how the choice of the parameter *k* will influence the performance of the algorithm and for this reason, its value must be carefully selected through *cross-validation* or similar techniques that can help identifying the optimal value for *k* on a given dataset.

4.3.3. Support vector machines

Support Vector Machine is one of the most famous and newest supervised learning algorithms, which is used primarily for classification problems in machine learning, but could also be useful to solve regression problems. The model complexity of an SVM is unaffected by the number of features encountered in the training data. For this reason, SVMs are well suited to deal with learning tasks where the number of features is large with respect to the number of training instances (Kotsiantis et al., 2006).

The ultimate goal of the SVM algorithm is to create an *"hyperplane"*, namely the best line or boundary that can separate n-dimensional space into classes (N-numbers of features) to distinctly classify the data points. Specifically, SVM selects the extreme points that are useful to define these decision boundaries, and these points work as a sort of support vector, from which the name of the algorithm is derived.

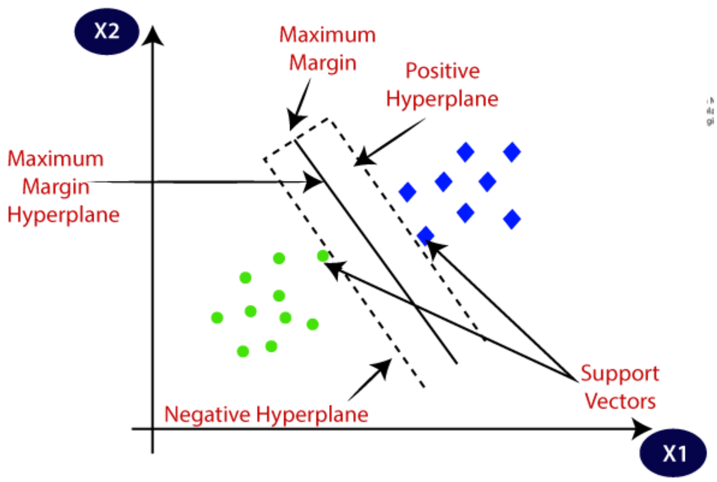


Figure 10: SVM Hyperplanes

As it could be seen in figure 10, the points constituting the boundaries are called support vectors, and the middle of the margin is the optimal separating hyperplane. Since the hyperplane is used to separate the data points, there are many possible boundaries that could be chosen. The objective is to find a hyperplane that has the maximum margin, namely the maximum distance between data points of both classes. Indeed, choosing the maximum distance allows future data points to be classified with more confidence.

Nevertheless, most real-world problems involve non-separable data for which no hyperplane exists that successfully separates the positive from the negative instances in the training set. To deal with the inseparability problem a possible solution is to map the data onto a higher-dimensional space and define a separating hyperplane there. The aforementioned higher-dimensional space is called the *feature space*, and by choosing it with an appropriate dimensionality, any training set can be made separable. On the other hand, the *input space* is occupied by the original training instances, and the latter are not linearly separable. Indeed, according to Kostiantis et al. (2006), a linear separation in feature space corresponds to a non-linear separation in the original input space, and it is noteworthy that this input space has a dimension equal to the number of features involved in the classification. For this reason, SVMs adopt Kernel functions in order to connect the values of the features from the input to the feature space, and in this way all the operations on values can be performed directly in the feature space, without the need of a complex mapping (Scholkopf et al., 1998).

Once the process is completed and a hyperplane has been developed, the Kernel function is used to map new points into the feature space for classification. Typically, more than one Kernel function can be adopted to map points from the original input space to the feature one: the selection of the most appropriate Kernel function is very important, since the latter defines the feature space in which the training set instances will be classified. For this reason, the choice of the function to be adopted is taken based on a cross-validation, after different functions are used to create models.

However, this is the reason why a limitation of SVMs is the low speed of the training. As long as the Kernel function is legitimate, a SVM will operate correctly even if the designer does not exactly know which features of the training data are being used in the Kernel-induced feature space. Even considering the low speed of training, the classification output obtained thanks to the SVM technique, has very good performances compared to other algorithms. Indeed, the training optimization problem of the SVM necessarily reaches a global minimum, and avoids ending in a local minimum, which may happen with other search algorithms such as neural networks. But, since the SVM methods are binary, in case of multiclass problem, before addressing the global problem, it has to be reduced to a set of multiple binary classification problems (Kotsiantis et al., 2006).

4.3.4. Logic-based algorithms

According to the different typologies of algorithms identified and classified by Kostiantis and colleagues (2006), the main group of logic (symbolic) learning method is represented by *Decision trees.*

Decision Trees

Decision trees belong to a class of supervised machine learning algorithms, which are used in both classification (predicts discrete outcome) and regression (predicts continuous numeric outcomes) predictive modelling. Decision trees are the simplest way of classifying *"objects"* in a finite number of classes. A decision tree is a system with *N* input variables and *M* output variables: it classifies instances by sorting them based on feature values. Input variables (attributes) are derived from the observation of the environment. The output variables, on the other hand, identify the decision/ action to be taken. Basically, a decision trees is made of three main elements: *nodes, branches* and *leaves*. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. (Murthy, 1998).

Nodes are tests performed on attributes associated to the data; the root node is the node at the top of the tree, and it is the only node without any incoming branch. Branches represent the values of the attributes and the leaves represent the class labels (Sá et al., 2016). Leaves are also known as terminal nodes or decision nodes. The figure 11 represents in a detailed way the structure of a decision tree:

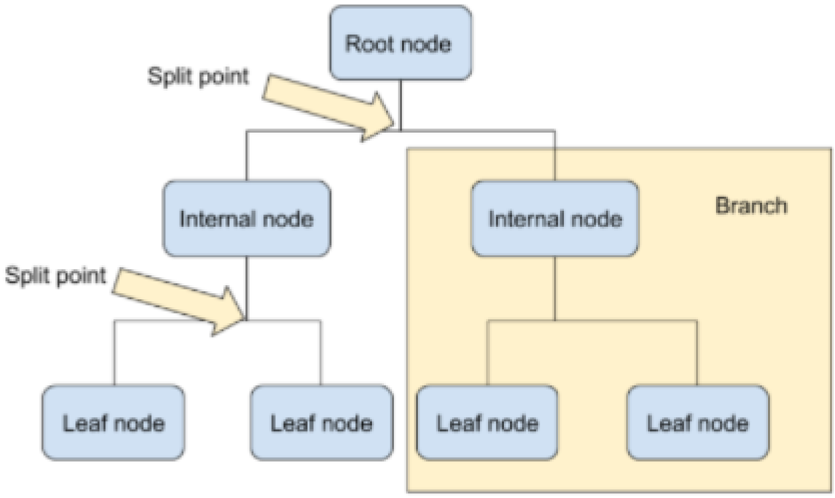


Figure 11: Decision trees

According to the figure number 11, it is possible to notice that the decision-making process is represented with an inverted logic tree where each node is a conditional function. Each node represents a condition (test) on a particular environment property (variable) and has two or more downward branches in operation. The process consists of a sequence of tests. It always starts from the root node, the parent node located higher up in the structure, and then goes down. Depending on the values found in each node, the flow takes one direction or another and proceeds progressively downwards.

From this description, since the resemblance of the decision tree allows to visualize a precise sequence of steps, researches agree that this chart could help to define a course of action, starting from a decision *("node")* and resulting in different branches that represent a possible outcome or reaction, so to figure out and understand in a very effective way all the potential outcomes that a decision could lead to. The trees also allow practitioners to visualize every potential option and weigh each precise course of action against the risks and rewards that each of these options can bring. Indeed, each end-result has an assigned risk and reward weight.

The development of a decision tree encompasses two phases: building according to the right splitting rules and classification.

The building procedure starts from a training set for the initial growth phase of the model, then a pruning phase follows. More in depth, in the first phase of *"building"* the data set devoted to the training-step is partitioned recursively until all the records in the partition have the same class, and for each class a new *"node"* is added to the decision-tree. The idea provided by Picasso Minicourse is to divide data into disjoint subsets according to some splitting rules, then repeat recursively for each subset and stop when leaves are (almost) *"pure"*. Indeed, the final aim of the building phase is to construct a *"perfect"* tree that accurately classifies every record from the training set.

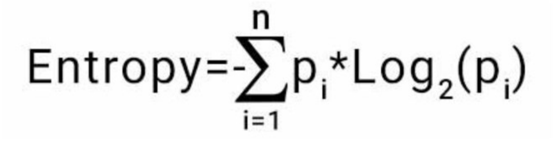
In this building phase, choosing the right splitting rules and so structure a tree according to some precise features or criteria is fundamental to obtain an effective decision model. Indeed, it is worthy to think that in a dataset of *N* attributes, deciding which are the attributes to put at the root and which are the ones to put at different levels of the decision-tree as internal nodes could be a very difficult issue.

To solve this difficult step, studies about decision tree charts agree to choose a rule that "*leads to greatest increase in purity"* (Picasso Minicourse), using some measures like *Entropy* and *Information Gain or Gini Index.*

Entropy and Information Gain

Entropy and Information Gain are two related concepts since the one of entropy plays an important role in calculating Information Gain.

In the Lyman works, entropy is not just the measure of disorder, or measure of purity. Basically, "*it is the measurement of the impurity or randomness in the data points",* calculated between the range of 0 and 1. So, entropy is a measure of the order of records that are considered for the construction of decision trees. A high entropy value expresses the *"disorder"* that characterizes the space of records, or greater difficulty in assigning each record to its class, based on the attributes that characterize the class itself. Higher the entropy, the less information we have on the attribute class. The formulation of Entropy is provided below:



$$\text{Entropy}=-\sum_{i=1}^{n} p_i * \text{Log}_2(p_i)$$

Entropy Formula

where p_i is the probability of an instance to be assigned to a specific class, and n is equal to 2 in case of a binary classification.

The Information Gain index is applied to quantify which feature provides maximal information about the classification based on the notion of entropy, by computing the difference between the entropy of the distribution before the split and the entropy of the distribution after the split. The greater this quantity, the higher the decrease of the entropy after have partitioned the data with the attribute. Therefore, information gain measures if the entropy of the system is lowered after splitting.

Information Gain (feature) = Entropy (Dataset) - Entropy (feature)

In conclusion, a criterion for choosing the nodes of a classification tree consists in choosing from time to time the attribute that gives a greater reduction of Entropy or that similarly maximizes the Information Gain.

Gini Index

Another important splitting index among the ones available is the Gini one, also called Gini impurity, which measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. It is calculated as 1 minus the sum of squared probabilities of each class. The index works on categorical variables and provides outputs either be *"successful"* or *"failure",* conducting binary splitting only. It is computed as:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Since the index conducts only binary splitting, its degree varies between 0 and 1, where: 0 denotes that all elements belong to a certain class or if only one class exists, and 1 denotes that the elements are randomly distributed across various classes.

Moving on, considering the building phase of the decision trees, at the end of the growth stage, usually complete trees are obtained, consisting of a very large number of rules. To have a tree which is effective in classifying records and providing rules, `it is necessary to *"thin out"* the redundancy of the tree itself, removing the less significant branches (*pruning*). Indeed, according to Rokach and Maimon (2006), the stop of the growth phase happens when certain conditions are met: for example, when all the observations belonging to the training set instances are assigned to a class, or when the maximum number of levels of the tree is achieved, or even when the best splitting criteria is not higher than a specified threshold. Then the second pruning stage should begin.

Indeed, in the greatest majority of the cases, the pruning phase is implemented after the building one, removing some of the branches once the tree is generated (Kotsiantis, 2013). In reality, the pruning stage could also be deployed before the initial phase, and, in this case, it is called pre-pruning. The latter consists in posing a condition to terminate some of the branches when the tree is built. However, the pruning phase implemented after the generation of the tree is generally faster and more accurate (Patil et al., 2010).

Moreover, it is important to mention, that this following pruning phase is not only useful to decrease the complexity of the tree (eliminating all the irrelevant features), but also to avoid the risk of overfitting. This phenomenon occurs when the algorithm adapts (fitting) too well (over) to training data losing its generality. So, apparently the model looks perfect on training data, but when applying it to testing data, many errors occur (for example in term of accuracy).

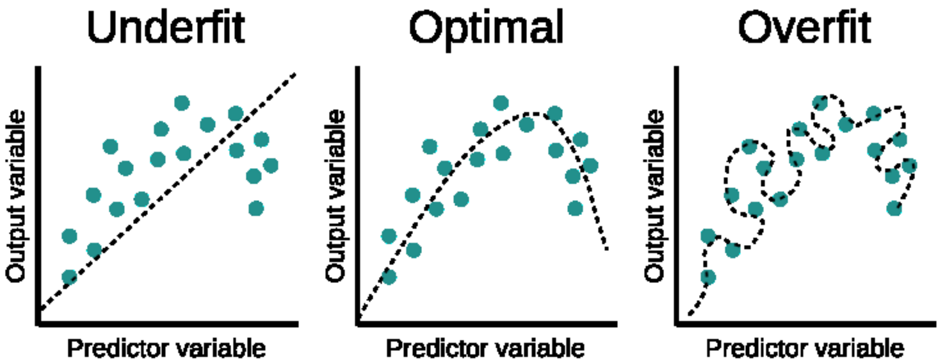


Figure 12: Overfitting problem

Another strategy that could be implemented to avoid the problem of overfitting is data pre-processing. It is performed before building the classification tree and, for this reason, does not affect directly the decision tree, but it acts on the training set.

Basically, the objective of data pre-processing is to select the most relevant features for building a simpler decision tree (Kotsiantis, 2013).

After the building phase is completed, the second one begins, namely a classification procedure: in this stage, the data of the test-set are classified according to the tree created and refined before. The classification steps make sure that to each observation, based on the features and the tests made by each node on the selected features, a class label is assigned. The class label is assigned to an observation when, starting from the root node and following the path of the decision tree, the leaf node is encountered (Kotsiantis, 2013).

When referring to trees model, according to Kostiantis (2013), the last topic noteworthy mentioning is the tree size, crucial to obtain a good accuracy in the output of the model. Indeed, trees must be big enough to fit training data, so that *"true"* patterns are fully captured. But, on the other way around as previously said, trees that are too big may overfit, thus capturing noise or spurious patterns in the data. This could cause a decrease in the accuracy of the model and reduce the reliability in classifying the observations. Typically, the most significant issue to overcome consists in predicting the best tree size from the training error, so in the test phase some problems may arise. Below in figure 13, a graph representing the relation between the tree size and the error rate is reported:

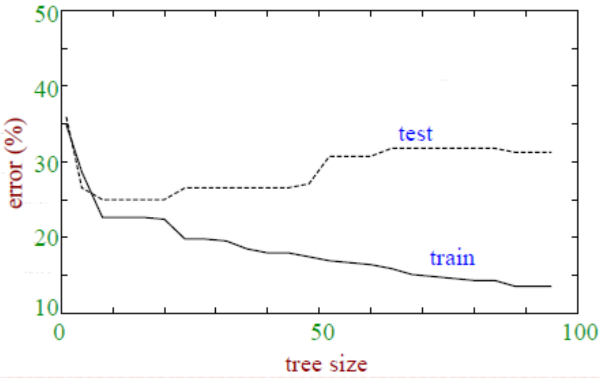


Figure 13: Relation between tree size and error rate

Finally, having seen how the model is shaped and works, and which are its main variables, it is now possible to sum up this algorithm by tracing its benefits and drawbacks. First of all, according to its simple representation, it is considered very easy to interpret and understand. It could also work on numeric and categorical data simultaneously, without requiring a huge amount of these data. However, the algorithm performs well on large datasets as well as on small groups of data.

Also, several drawbacks could be associated to these supervised trees. Indeed, typically, some greedy algorithms could suffer from bias and variance, and for this reason, cannot guarantee to return the globally optimal decision tree. This result can be mitigated by training multiple trees (ensemble of decision trees) where the features and instances are randomly sampled with replacement.

Ensemble of decision trees

Ensemble methods combine several decision trees to produce better predictive performances compared to a single decision tree. This model is referred to the idea proposed by Rokach and Maimon (2006), who stated that by combining a set of models solving the same task, a better global model, with more reliable predictions, can be developed. Indeed, the main principle behind the ensemble model is that a group of weak learners are put together to form a stronger learner, with more trustworthy predictions. Different techniques to perform ensemble decision trees are available, the ones worthy to mention are *bagging* and *boosting*.

Bagging

Bootstrap aggregation, also called bagging (from bootstrap aggregating), is applied when small variations in data might result in unstable decision trees, and there is the need to reduce this *"variance"*. Indeed, this ensemble algorithm is designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression, reducing variance and helping to avoid overfitting.

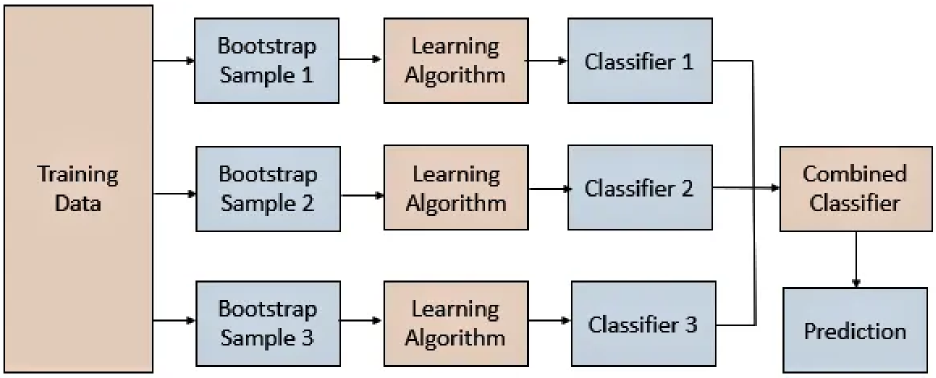


Figure 14: Bagging methods

Bagging method is characterised by two phases: aggregation and bootstrapping. Bootstrapping is a sampling method, where a random sample of data in a training set is chosen with replacement, so an individual data point could be selected more than once. As a result, a value or instance could be repeated twice or more in a sample. Then, the learning algorithm is trained and run on the samples selected in parallel, as shown in the figure 14, using weak or base learners. Finally, the aggregation phase begins and according to the task, if regression or classification, an average or a majority of the predictions are taken to compute a more accurate estimate.

As a note, the *"random forest"* algorithm is considered an extension of the bagging method, using both bagging and feature randomness to create an uncorrelated forest of decision trees.

Boosting

Boosting is an alternative learning method with respect to bagging, that combines weak learners, also called base learners, into a strong rule to minimize training errors. The model works by selecting random sample of data, then these sample are trained sequentially. Indeed, the main difference between this algorithm and bagging is characterised by the way in which they are trained: in boosting, they learn sequentially, differently from the *"parallel"* training of bagging. More in detail, this sequentially procedure implies that a series of models are built and with each new model iteration, the weights of the misclassified data in the previous model are increased. So, each time that the learning algorithm is run out, it provides a new weak prediction rule, and, after many iterations, the boosting technique combines these weak rules into a unique strong prediction rule.

Rule – Based Classifiers

According to Quinlan (1993), decision trees can be translated into a set of rules by creating a separate rule for each path: for each sequence of the tree algorithm, going from the root node to a leaf node, it is possible to define a distinct rule. However, rules can also be directly inferred from training data using different rule-based algorithms. The process through which rule-based classifiers work is an iterative one, indeed each time the algorithm tries to find the pattern that explains as many instances as possible, so that the number of excluded instances is high. To achieve this objective, as reported by Dutton and Conroy, 1997, the search starts with general rules consisting of one attribute each, and then it continues with more complex rules as the number of remaining instances decreases. Kostiantis and colleagues (2006) stated that the goal is to construct the smallest rule-set, consistent with the training data. Indeed, a large number of learned rules is usually the consequence that the learning algorithm is trying to *"remember"* the data training, instead of discovering the assumptions that rule it.

There are two main kinds of rule-based classifiers: *separate-and-conquer* approach and a *divide-and-conquer* approach; the main difference among them is that separate-and-conquer approach does not impose conditions separating an attribute on its values, but rather generates rules with one attribute-value pair at a time (Lindgren, 2004). Moreover, the rules introduced in the learner have to be both reliable and with a high predicting power. For this reason, they are subject to an evaluation through a function that tests their quality, statistical significance and possibility of specialization (Dutton and Conroy, 1997).

Inductive logic programming

When Kostiantis and his colleagues proposed the algorithm classification (2003), they defined as the main restriction of propositional learners their limited capacity to take into account available background knowledge. Instead, inductive logic programming (ILP) is a subfield of logic artificial intelligence which uses programming as a uniform representation for examples, background knowledge and hypothesis, when performing data analysis. Precisely, ILP classifiers use first order predicate logic: ILP system derives a hypothesised logic program which entails all the positive and none of the negative examples.

Positive examples + negative examples + background knowledge ⇒ hypothesis

ILP systems typically adopt the covering approach of rule induction systems. Indeed, these systems initially develop a clause in line to justify and explain a good number of positive observations, since they typically prefer clauses able to include as many instances as possible. Then, this first clause is added to the hypotheses built from background knowledge and new clauses are provided to explain all the other remaining positive observations. Typically, a search starts with a very general rule, with no conditions. Then, conditions are added to the first clause until it only covers (explains) positive observations. When a clause explains only the positive observations, we can say that the clause is consistent (Kotsiantis et al., 2006).

As said, inductive logic programming differs from most other forms of machine learning not only for its capacity to make use of logically encoded background knowledge, but also for its use of an expressive representation language. For this reason, ILP systems have been employed for natural language processing with successful applications (Muggleton, 1999). Therefore, given the possibility to perform analysis in a more expressive language, these techniques need a higher computational capacity.

4.4. Steps for developing a classification model

4.4.1. Data Preparation

To develop a classification model, a set of observations, characterized by a certain number of attributes, with known values, and a class label, are given as input to an algorithm. Obviously, these data provided to the classification algorithm play a central role, this is why it is fundamental to prepare them in a proper way, so to achieve the best outcome possible. Starting from the available samples and leveraging on data provided about the values of their attributes, the algorithm can infer a model able to predict the class label of future observations.

Data preparation is needed as the data collected from the input sources could have anomalies or bias, which might be then reflected in the model developed by the algorithm. To avoid these issues, three steps are performed on the raw-data collected: *data validation, data transformation* and *data size reduction* (Vercellis, 2009).

Data Validation

Data validation is a process that ensures the delivery of clean and clear data to the programs, applications and services using it. This phase checks for the integrity and validity of the data inputted to different software. Data validation ensures that the data complies with the requirements and quality benchmarks. Indeed, the quality of input data could be lowered if they are incomplete or if the observations are affected by any kind of noise.

Data are said to be incomplete if the observations have some missing values. In order to solve this issue, there are many techniques to be used, such as *elimination*, *inspection*, *identification* and *substitution*. Elimination means discard all the observation with missing or incomplete values for one or more attributes; inspection refers to the process whereby experts analyse observations with missing values one by one. At the end of the process the expert can suggest substitute values and this is a very time-consuming procedure. Instead, through identification the analysts substitute all the missing values with values that are not possible for the attributes considered (e.g., the value -1 could be used for attributes that only allows positive numerical values). The last method is substitution, which differs from both inspection and identification, since here all the automatic replacement techniques are considered. Such techniques imply the substitution of the missing values with the mean of the values calculated starting from the remaining observations or by using the maximum likelihood value, estimated using regression models or Bayesian methods in case the distribution of values is not symmetric.

As mentioned at the beginning, the data may not only be incomplete, but there could be a possible noise affecting the dataset, which must be handled. Noise happens when there are erroneous or anomalous values (i.e., outliers), or the values associated to an attribute are expressed in heterogenous measurement units or, in addition, when the data are collected with a process that induces errors. The first action to be done to manage the noise consists in eliminating the outliers from the dataset. The procedure to identify the outliers is based on looking at the distribution of the values associated to an attribute and relies on the assumption of gaussian distribution of the values of the attribute: if the value of one or more observations falls outside an appropriate interval centred around the mean value for that attribute, it means that value is an outlier. Moreover, in cases in which the actual distribution of data is not symmetric other methods can be applied, as for example identification of outlier through cluster analysis.

Data Transformation

Data transformation is the process of changing the format, structure, or values of data, in order to facilitate the learning phase and improve the accuracy and efficiency of learning algorithms. There are different data transformation techniques available, but the most adopted ones on input data are *Standardization* of the data and *feature extraction*. Among the standardization techniques, it is necessary to go into the detail of the min-max one, which will be necessary for a thorough understanding of the methodology adopted for the realization of this thesis.

Min-max standardization helps to normalize the data by scaling them between 0 and 1, and so, helps to understand the data easily. The standardization is obtained applying the following formula for each value observed:

$$x'_{ij} = \frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}} \times \left(x'_{max,j} - x'_{min,j}\right) + x'_{min,j}$$

Where the letter i refers to the i^{th} observation, while the letter j is used to indicate to indicate the j^{th} attribute. The values of $x'_{max,j}$and $x'_{min,j}$ set the extremis of the desired range. If $x'_{max,j}$=1 and $x'_{min,j}$=0, the equation can be re-written as follow:

$$x'_{ij} = \frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}}$$

Feature extraction works through transformations of the attributes already available in order to generate new ones. This process is particularly useful for turning the variables already available in more informative variables, which summarize additional relevant information.

Data Reduction

The size of the dataset used for the classification strongly impacts on the computation time needed to train the models. This is the reason why performing a data reduction procedure may be very beneficial to minimize the time required for training the model and perform the analysis more quickly. To do so, there are two different ways, namely acting on the number of observations considered with *data sampling*, or to reduce the number of attributes through *feature selection*.

Data sampling is a statistical technique which chooses and analyses a representative subset of data points, which will be characteristic of an entire group as a whole, to identify patterns and trends in the complete data set. In this way, it is possible to work with a manageable amount of data, while still producing accurate findings.

Feature selection works on the number of attributes, so the reduction in data involves values present in all the observations. Indeed, the selection underlines the less relevant attributes for the classification which can be eliminated from the original dataset, namely the ones that are redundant or irrelevant in the presence of another relevant features with which are strongly correlated. Three main feature selection methods are known: *Filter, Wrapped* and *Embedded selection*, based on how they combine the selection algorithm and the model building.

Filter methods select features from a dataset independently from any machine learning algorithm. These methods rely only on the characteristics of the variables, so features are filtered out of the data before learning begins. These methods are powerful, simple and help to quickly remove features: generally, they constitute the first step in any feature selection pipeline.

Wrapped Methods are an automatic feature reduction process requiring no human intervention. Indeed, these methods involves a *"search strategy"* selecting different subset of data and evaluating each of them based on the quality of the performance of the given algorithm. If an attribute does not increase the accuracy of the model, then it is considered eliminable by the selection performed with Wrapped method. Basically, it is an iterative process in which another model is created by selecting features until the subset of factors that gives the maximum accuracy of the prediction is found.

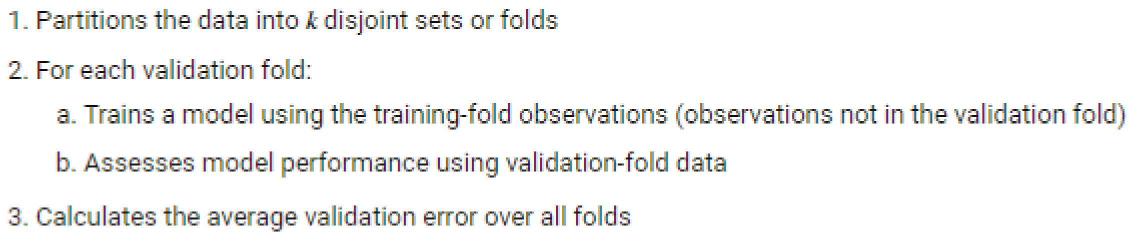
In *embedded methods* modelling algorithms are adopted. The latter use the coefficients of features to reduce them. Embedded methods are among the most sophisticated feature selection methods as it combines the advantages of Filter and Wrapped methods. For Embedded methods, the features have to be on the same scale.

4.4.2. Training phase

The dataset provided in input to the classification algorithm is usually divided into two subsets to firstly perform the training of the model and then its testing.

During the training phase, the so-called training dataset is fed to the model, so that it is configured adapting its parameters to the data received as input. Thanks to the training of the model, the classifier will learn to derive rules and patterns that allow the identification and association of a class label, and finally to newly provided observations during the testing phase. Before starting these second phase, it is important to understand the accuracy of the trained model, namely to understand whether it has good performances or not. This can be done relying on *cross-validation techniques,* which imply the splitting of the training set into a fixed number of mutually exclusive and equal-sized subsets. Every round, one of them is used as validation set and all the others are used to train the model. The output of this procedure is the evaluation of the error rate for each one of the validated subsets, so that the average of these error rates is the error rate of the classifier.

Starting from the computation of the error it is possible to estimate the accuracy of the trained models before the testing phase, so that starting from the training set it is possible to evaluate different algorithms and assess which one is the most accurate for the data considered. What is obtained at the end of the cross-validation is therefore used to perform the algorithm selection and identify the best classifiers. In the following figure, the main steps of the cross-validation procedure:



1. Partitions the data into *k* disjoint sets or folds
2. For each validation fold:
    a. Trains a model using the training-fold observations (observations not in the validation fold)
    b. Assesses model performance using validation-fold data
3. Calculates the average validation error over all folds

Figure 15: k-fold cross-validation algorithm. Source: https://www.mathworks.com/help/stats/select-data-and-validation-for-classification-problem.html

4.4.3. Testing phase

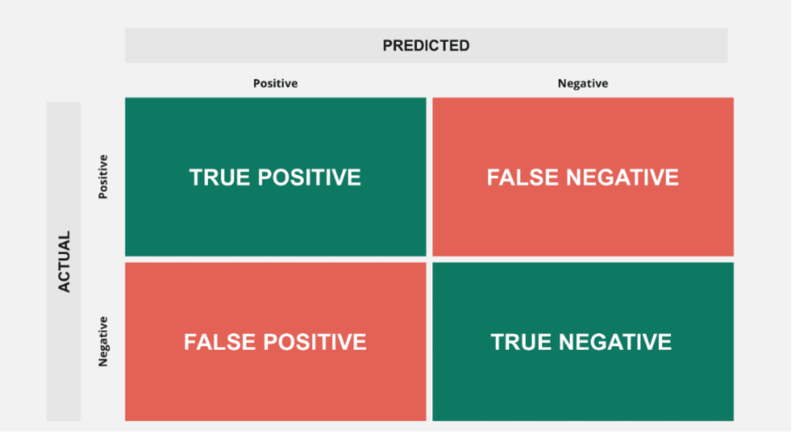
The last phase when building a classification model is the testing phase, and it consists in providing the testing dataset to the trained model in order to assess the behaviour of the classifier.

During this phase, the classifier previously trained, will be given as input unseen observations, and based on what it has learnt during the previous phases, it will be able to correctly handle and classify the new samples. In order to assess how much accurate the classifier is, different metrics are computed.

In the subsequent paragraphs, we will describe the most common which are *accuracy, precision, recall, F1 score.*

Confusion Matrix

A confusion matrix, also known as error matrix, is a specific table layout that allows visualization of the performances of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa. In the figure (16) a two-class confusion matrix is shown.



		PREDICTED	
		Positive	Negative
ACTUAL	Positive	TRUE POSITIVE	FALSE NEGATIVE
	Negative	FALSE POSITIVE	TRUE NEGATIVE

Figure 16: Confusion matrix

On the green diagonal we find the total number of right predictions made by the classifiers. This value is given by the sum of the *true positive*, namely those instances that were true and that the classifier has identified as true, TP and the *true negative*, namely those instances that were false and the classifier has correctly identified as false, TN. The formula to calculate the total number of right predictions is the one below:

$$Total\ number\ of\ right\ predictions = TP+TN$$

On the other hand, on the red diagonal we find the sum of the total instances wrongly classified. There are the *false negative* (FN), namely those instances that are true but the classifier identifies as false, and the *false positive* (FP), namely those instances that are false but were wrongly classified as positive. The total number of wrong predictions can be calculated following the formula below:

$$Total\ number\ of\ wrong\ predictions = FN+FP$$

In order to obtain the total number of instances tested, the equation shown below shall be applied:

$$Total\ number\ of\ tested\ instances = TP+TN+FN+FP$$

Accuracy

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points, namely the proportion of correct predictions in all predictions made. More formally, it provides the percentage of right predictions over the total number of tested instances, as shown in the equation below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Total\ number\ of\ right\ predictions}{Tota\ number\ of\ tested\ instances}$$

Often, accuracy is used along with *precision* and *recall*, which are other metrics that use various ratios of true/false positives/negatives. Together, these metrics provide a detailed overview of how the algorithm is classifying instances.

Precision

As a measure, Precision is defined as the proportion of correct positive predictions of all cases classified as positive. For this reason, it is a useful metrics for assessing how good the classifier is in detecting the true positive instances and it is also suitable for capturing the costs of false-positive assessments.

$$Precision = \frac{TP}{TP + FP}$$

Recall

It provides the percentage of actual positive instances that have been correctly identified. It is computed as the ratio between true positive instances and the sum of true positive instances and false negative ones, as shown in the formula below:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

The *F-score*, also called the F1-score, is a measure of a model's accuracy on a dataset. It is adopted to evaluate binary classification systems, which classify examples into *"positive"* or *"negative"* and it is useful when it is more important to detach false negatives and false positives.

This metric is a way of combining the precision and recall of the model, and so it is defined as the harmonic mean of the model's precision and recall. It is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

It is possible to adjust the F-score to give more importance to precision over recall, or vice-versa. Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score. It can be calculated following the equation below:

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}$$

ROC Curve

A *ROC curve* (Receiver Operating Characteristic Curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, true positive rate (TPR) and false positive rate (FPR). In the figure (17) a ROC curve is shown.

On the X-axis of the ROC curve, the false positive rates are found, also computed as *100-specificity*. The false positive rate is the ratio between the false positive (FP) instances and the sum of false positive and true negative instances (FP+TN). Specificity is a measure of how well a classifier can identify true negative instances. On the Y-axis of the ROC curve, we find the recall, which is also known as true positive rate or sensitivity.

The ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold, more items are classified as positive, thus increasing both False Positives and True Positives. So, if a threshold has to be exceeded in order to be identified as belonging to the class 1, all the instances that have a value higher than the threshold will be classified as 1, otherwise they will be assigned to the class 0.

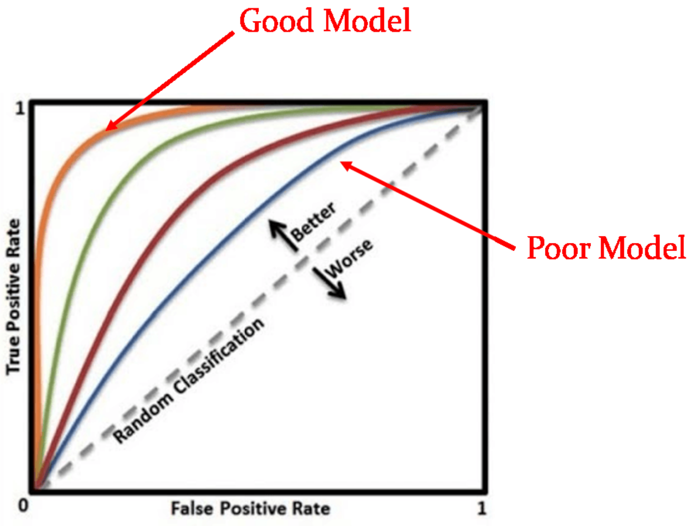


Figure 17: ROC Curve

There are two extreme cases. The first one is when the classification threshold is set equal to 0, all the instances will be classified as belonging to the class 1 and the true positive rate and false positive rate will be equal to 1. The second case is when the threshold is set at 1, so zero instances will belong to the class 1, and both the rate of true positive and false positive will be equal to 0, because, given the high threshold, the classifier has not predicted any positive instances. The presence of these two extreme cases implies the passage of the ROC curve from the points (0;0) and (1;1).

To compute the points in a ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information for us, called AUC. AUC stands for *"Area under the ROC Curve",* and it measures the entire two-dimensional area underneath the ROC curve (think integral calculus) from (0,0) to (1,1) and provides an aggregate measure of performances across all possible classification thresholds.

Chapter 5 – Dataset and Predictors description

5.1. Data collection

Regarding the following phase, we employed a database previously collected (Annunziata e Aversa, 2020) which includes 352 Kickstarter campaigns launched and concluded between 2016-2017, with a duration of at least 7 days. Among them, 195 are unsuccessful, namely failed to reach the target by the end of the campaign, and, on the contrary, the remaining 157 managed to reach and/or overcome the goal by the deadline.

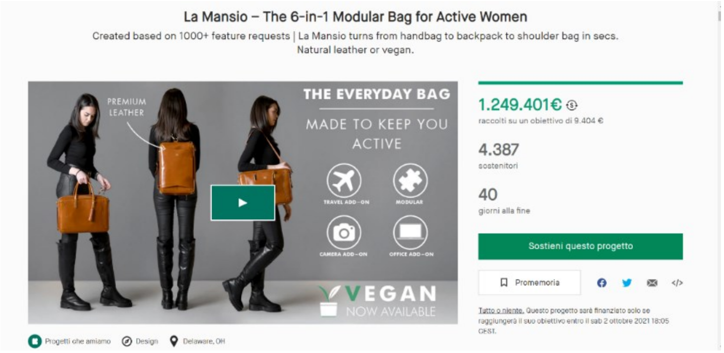
Firstly, we explain into details the study context from which the data were extracted, in order to have a clear overview of the platform involved and its main features. Then, all the variables employed in our analysis are presented, with the main motivations behind their choices.

5.1.1. The study context

To carry out the analysis, we used data gathered from Kickstarter[16], that as said in the Literature Review paragraph, is one of the world's biggest online reward-based crowdfunding platforms (to read further information regarding the mentioned platform, go to *paragraph 2.2.3.* Crowdfunding models).

When landing on a campaign's webpage, a lot of information is presented.

[16] La Mansio – The 6-in-1 Modular Bag for Active Women by La Mansio — Kickstarter



La Mansio – The 6-in-1 Modular Bag for Active Women

Created based on 1000+ feature requests | La Mansio turns from handbag to backpack to shoulder bag in secs. Natural leather or vegan.

1.249.401€

raccolti su un obiettivo di 9.404 €

4.387

sostenitori

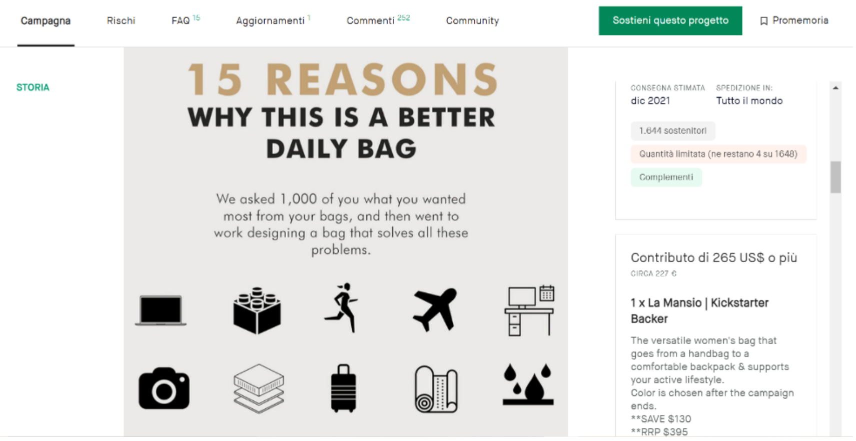
40

giorni alla fine

Tutto o niente. Questo progetto sarà finanziato solo se raggiungerà il suo obiettivo entro il sab 2 ottobre 2021 18:05 CEST.

Figure 18: Kickstarter Homepage

Firstly, the campaign title appears at the top of the page; instead, in the right part, it is possible to see the *target*, the *pledge reached*, the *number of backers* and *how long it takes to expire the campaign*. In this same part of the page the *link* that leads any lenders to make their donation is also displayed.



We asked 1,000 of you what you wanted most from your bags, and then went to work designing a bag that solves all these problems.

CONSEGNA STIMATA: dic 2021

SPEDIZIONE IN: Tutto il mondo

1.644 sostenitori

Quantità limitata (ne restano 4 su 1648)

Complementi

Contributo di 265 US$ o più

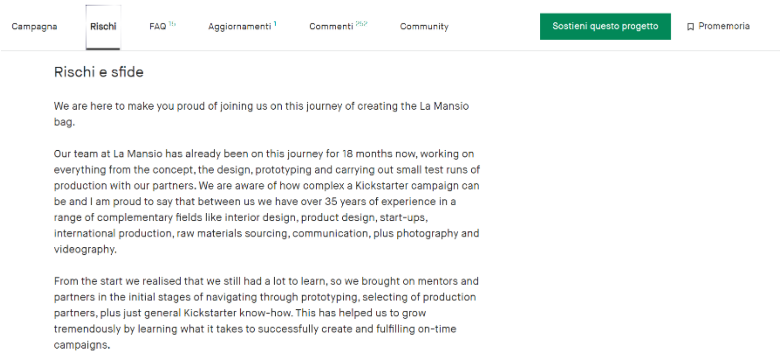
CIRCA 227 €

1 x La Mansio | Kickstarter Backer

The versatile women's bag that goes from a handbag to a comfortable backpack & supports your active lifestyle.
Color is chosen after the campaign ends.
**SAVE $130
**RRP $395

Figure 19: Kickstarter campaign tab

Scrolling through the page again, several tabs are displayed. The first one is the *Campaign Tab*, where the potential backer can visualize the story section. In the latter, the fundraiser tells his/her own story and explains the idea into details, enriched with images and videos to fully show the intended project's outcome.



Rischi e sfide

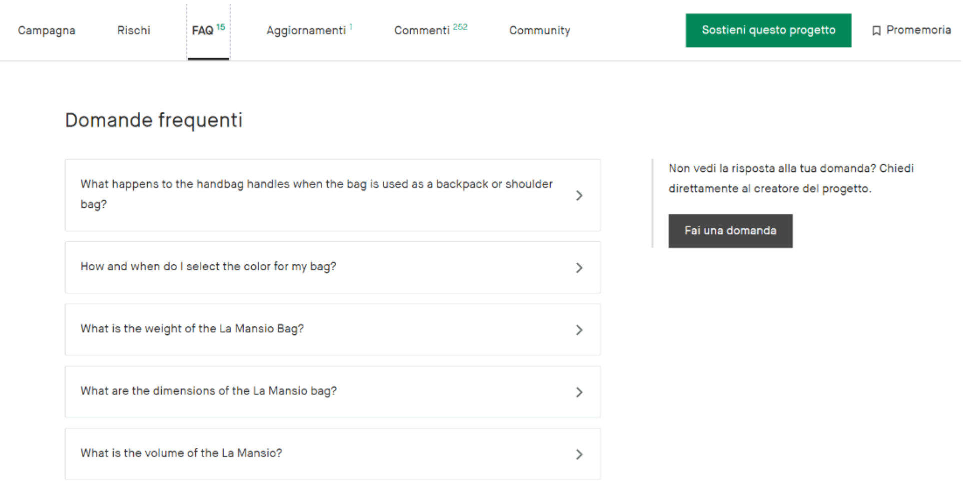
We are here to make you proud of joining us on this journey of creating the La Mansio bag.

Our team at La Mansio has already been on this journey for 18 months now, working on everything from the concept, the design, prototyping and carrying out small test runs of production with our partners. We are aware of how complex a Kickstarter campaign can be and I am proud to say that between us we have over 35 years of experience in a range of complementary fields like interior design, product design, start-ups, international production, raw materials sourcing, communication, plus photography and videography.

From the start we realised that we still had a lot to learn, so we brought on mentors and partners in the initial stages of navigating through prototyping, selecting of production partners, plus just general Kickstarter know-how. This has helped us to grow tremendously by learning what it takes to successfully create and fulfilling on-time campaigns.

Figure 20: Kickstarter campaign risk tab

Then the *Risk Tab* follows in the page, in which the fundraiser explains all the problems and challenges that could prevent the idea from becoming a successful one.



Domande frequenti

What happens to the handbag handles when the bag is used as a backpack or shoulder bag?

How and when do I select the color for my bag?

What is the weight of the La Mansio Bag?

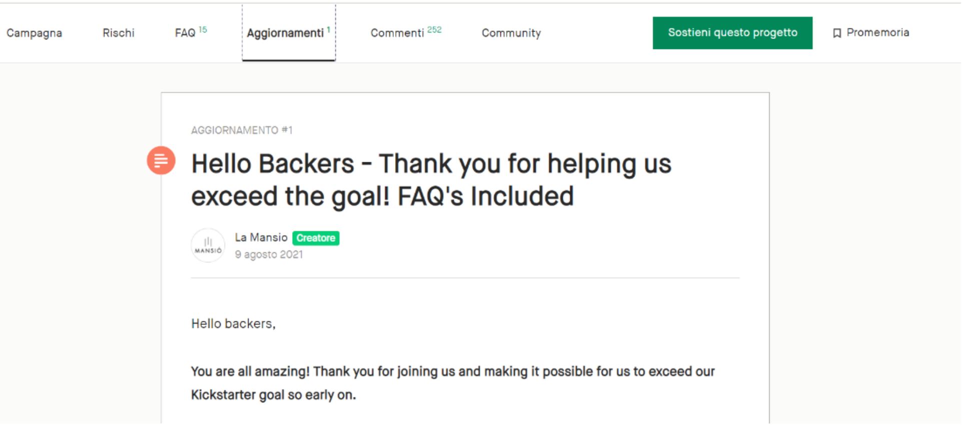
What are the dimensions of the La Mansio bag?

What is the volume of the La Mansio?

Non vedi la risposta alla tua domanda? Chiedi direttamente al creatore del progetto.

Figure 21: Kickstarter campaign Frequently Asked Questions tab

Going on, *Frequently Asked Questions* section is presented, which shows answers related to the shipment of the reward or potential delays.



AGGIORNAMENTO #1

Hello Backers - Thank you for helping us exceed the goal! FAQ's Included

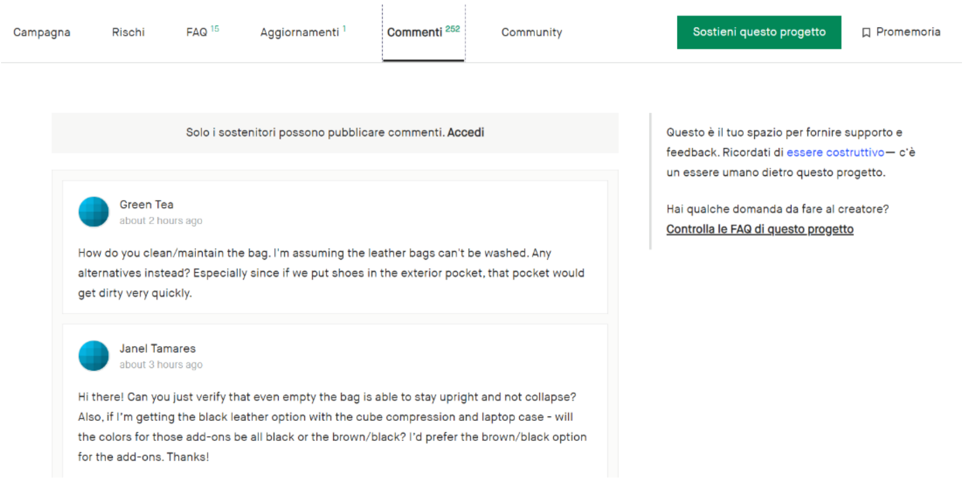
La Mansio Creatore
9 agosto 2021

Hello backers,

You are all amazing! Thank you for joining us and making it possible for us to exceed our Kickstarter goal so early on.

Figure 22: Kickstarter campaign Updates tab

Next to this section *Updates Tab* is displayed. Updates are posted by the fundraiser at the beginning and during the campaign duration, to provide useful insights and information regarding the project.



Solo i sostenitori possono pubblicare commenti. Accedi

Green Tea
about 2 hours ago

How do you clean/maintain the bag. I'm assuming the leather bags can't be washed. Any alternatives instead? Especially since if we put shoes in the exterior pocket, that pocket would get dirty very quickly.

Janel Tamares
about 3 hours ago

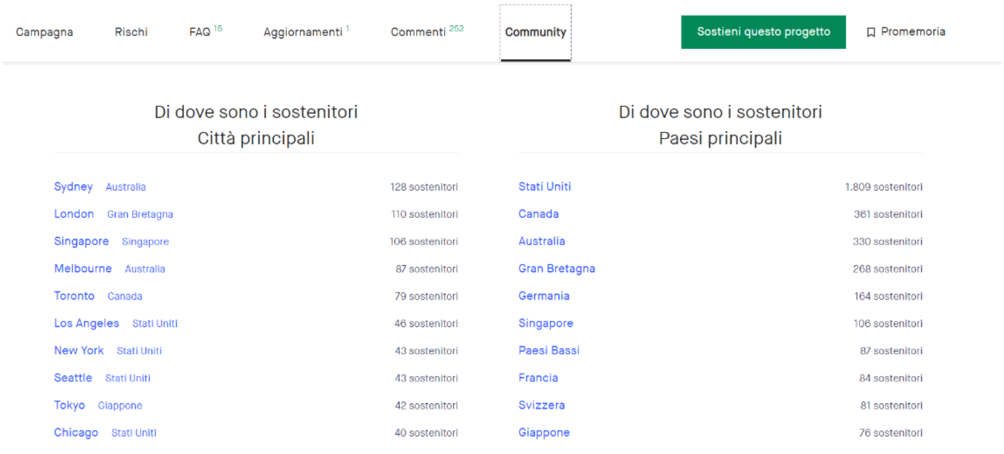
Hi there! Can you just verify that even empty the bag is able to stay upright and not collapse? Also, if I'm getting the black leather option with the cube compression and laptop case - will the colors for those add-ons be all black or the brown/black? I'd prefer the brown/black option for the add-ons. Thanks!

Questo è il tuo spazio per fornire supporto e feedback. Ricordati di essere costruttivo— c'è un essere umano dietro questo progetto.

Hai qualche domanda da fare al creatore? Controlla le FAQ di questo progetto

Figure 23: Kickstarter campaign Comments tab

Moving on, the *Comments Tab* is shown. Comments can be written by the project creator and everyone interested in the project profile during the campaign, and can be in the form of compliments, congratulations, questions and doubts as well as feedbacks and suggestions.



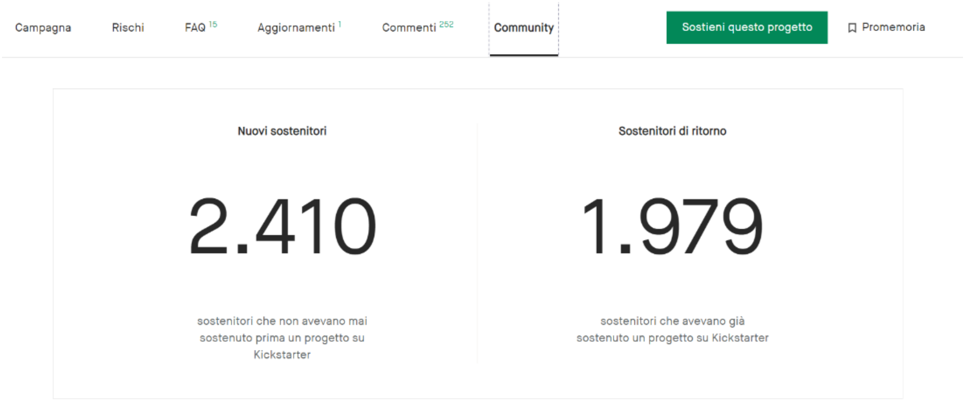
Di dove sono i sostenitori
Città principali

Sydney	Australia	128 sostenitori
London	Gran Bretagna	110 sostenitori
Singapore	Singapore	106 sostenitori
Melbourne	Australia	87 sostenitori
Toronto	Canada	79 sostenitori
Los Angeles	Stati Uniti	46 sostenitori
New York	Stati Uniti	43 sostenitori
Seattle	Stati Uniti	43 sostenitori
Tokyo	Giappone	42 sostenitori
Chicago	Stati Uniti	40 sostenitori

Di dove sono i sostenitori
Paesi principali

Stati Uniti	1.809 sostenitori
Canada	361 sostenitori
Australia	330 sostenitori
Gran Bretagna	268 sostenitori
Germania	164 sostenitori
Singapore	106 sostenitori
Paesi Bassi	87 sostenitori
Francia	84 sostenitori
Svizzera	81 sostenitori
Giappone	76 sostenitori

Figure 24: Kickstarter campaign Community tab, 1



Nuovi sostenitori

2.410

sostenitori che non avevano mai sostenuto prima un progetto su Kickstarter

Sostenitori di ritorno

1.979

sostenitori che avevano già sostenuto un progetto su Kickstarter

Figure 25: Kickstarter campaign Community tab, 2

The last section displayed is the *Community* one, which allows fundraiser and potential backers to visualize the cities and countries where backers come from, and how many new and returning backers are supporting the project.

5.2. Predictors

For each project extracted (352 campaigns) we considered in our analysis two types of variables: *static* and *dynamic variables*. The choice of the predictors has been based on past analysis of the Master Thesis *"Predicting the success of crowdfunding campaigns: a machine learning approach"* (Annunziata and Aversa, 2020) and on the literature reviews previously discussed in *Chapter 2*, since all of these features have been taken into consideration by scholars, during the last years, as important determinants of the crowdfunding campaign outcomes.

In the next paragraphs we dig deeper into a detailed introduction for each set of variables, both statics and dynamics.

5.2.1. Static variables

We define static variables those that are available from the launch of the campaign and that remain constant for its whole duration. In our database we have considered 20 static features: 11 *campaigns-related features*, namely those variables associated to the campaign characteristics, and 9 *linguistic variables*, which are directly derived from the textual description of the projects. Below, we will deepen each static feature considered.

Campaigns-related features

Among the 11 factors that belong to this category, the first variable that we will present is the one related to the category of the project that the platform Kickstarter associates to each campaign, "*s_category*", which is a categorical variable characterised by a finite number of different categories: Art, Comics, Crafts, Dance, Design, Fashion, Film and Video, Food, Games, Journalism, Music, Photography, Publishing, Technology and Theatre.

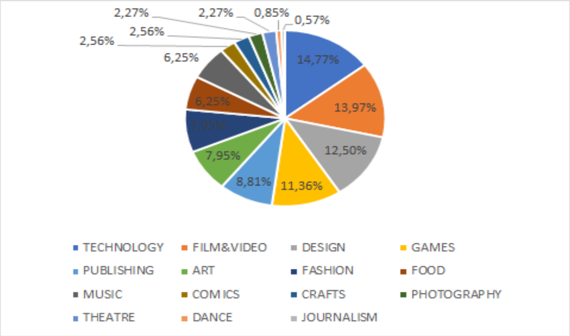


Figure 26: Chart of the project categories

According to the pie chart (26), we can state that our data sample covers all the categories available in the crowdfunding platform, with a majority of projects that belong to Technology (14,77%), Film & Video (13.07%), Design (12.50%) and Games (11.36%).

Category	***Number of occurrences***	***% on Total observed campaigns***
Art	28	7,95%
Comics	9	2,56%
Crafts	9	2,56%
Dance	3	0,85%
Design	44	12,50%
Fashion	28	7,95%
Film & Video	46	13,07%
Food	22	6,25%
Games	40	11,36%
Journalism	2	0,57%
Music	22	6,25%
Photography	8	2,27%
Publishing	31	8,81%
Technology	52	14,77%

Theatre	8	2,27%

Table 16: Project categories occurrency

In table (16), we display a first column representing the number of occurrences and a second one constituted by the number of observed campaigns for each category on the total number of campaigns.

"s_category" is the only categorical factor, instead the others can be described as numerical. One of these is the variable "*s_backed",* which is related to the backing experience of the fundraiser. Indeed, this variable originally represented the number of projects the fundraiser had already backed in the past, then we transformed it into a dummy variable equal to 1 if the fundraiser has backed at least one crowdfunding project on Kickstarter, 0 otherwise.

The variable "*s_country",* related to the project location, follows the same procedure: originally it was a categorial variable characterised by the name of the country in which the project was located. Then, since the crowdfunding platform of reference, i.e., Kickstarter, is US-based we transformed the categorical variable into a dummy one, with value equal to 1 if the project location is U.S., 0 otherwise.

The variable *"s_created"* codes the number of projects that the campaign's creator has launched so far, and assumes the value of the right quantity of campaigns already created.

The fifth variable presented is the duration set by the fundraiser, called *"s_duration"*. Its value identifies the lifecycle of the project on the platform, that is the number of days within which the target goal must be reached. The feature assumes the value of the days necessary to conclude the campaign.

Subsequently, we considered the gender of the fundraiser through the dummy variable *"s_gender"* with a value equal to 1 when fundraiser is female, 0 when the fundraiser is male. This variable was coded and valued by analysing the fundraiser's first name.

Moreover, the number of pictures and videos, that are showed and featured in the project description, are taken into consideration: *"s_picture"* and *"s_video"*. Both these variables assume the value of the exact number of pictures and videos displayed in the description section.

"s_serial" is a dummy which accounts for the internal social capital of the creator. The variable is equal to 1 if the fundraiser is considered *"serial"*, so if he or she could be defined as an *"entrepreneur who repeatedly turn to crowdfunding to finance their projects"* (Butticè and colleagues, 2017).

Additionally, the variable *"s_target"* identifies the goal amount set by the fundraiser expressed in US dollars. Since in our sample, some campaigns analysed referred to other geographical locations, such as England, the standard conversion method from local currency (such as pound) to dollar was used.

Finally, we defined the dummy *"s_team"*, which is 0 when the fundraiser is an individual, 1 when the project is created by a team. Since in Kickstarter it is not specified whether the fundraiser is an individual or a group, we considered a project as created by a team when the *"About the creator"* part involves at least one collaborator, or when this section refers to a group of individuals working on the project.

In the table 17 we summarize all the 11 campaign-related features previously analysed.

Campaign-related variable	*Description*
s_category	Categorical variables, it describes the Kickstarter category to which the project belongs.
s_backed	Dummy variable, equal to 1 if the fundraiser has already backed a project in the past; 0 otherwise.
s_country	Dummy variable, equal to 1 if the project is located in the United States; 0 otherwise.
s_created	Variable representing the number of projects already created by the fundraiser.
s_duration	Variable representing the campaign duration defined at the beginning by the fundraiser.
s_gender	Dummy variable, equal to 1 if the fundraiser is female; 0 if fundraiser is male.
s_picture	Variable representing the number of pictures displayed in the project description.
s_serial	Dummy variable, equal to 1 if the fundraiser is a serial one; 0 otherwise.
s_target	Variable representing the target goal set defined at the beginning by the fundraiser.
s_team	Dummy variable, equal to 1 if the fundraiser is team; 0 if the fundraiser is individual.
s_video	Variable representing the number of videos displayed in the project description.

Table 17: Campaign-related variables

Linguistic features

Different variables related to the text description and analytics of the campaign should be presented. Among all the 9 features, six linguistic ones have been extracted by Annunziata and Aversa (2020), using a computerized text analysis software known as "*Linguistic Inquiry and Word Count*". [17]As specified in their Dissertation, the two-thesis workers performed a linguistic analysis of the *title*, *blurb* and *description* of the campaign as a whole, with the aim to also include the fundraiser's written elements.

Precisely, the LIWC variables reported in their Master Thesis are: "*s_wordcount*" which counts the number of words within the text, "*s_tone*" that measures how positive the tone is according to the words used, "*s_percept*" which values the presence of perceptual words, such as *see*, *hear*, and *feel*, "*s_risk*" that measures the presence of risk-related nouns, "*s_money*" related to the usage of money-related words, and finally *"s_I"* counting the presence of the pronouns *"I"*.

Moving on to the other three variables, "*s_agentic"* is related to the extent to which the fundraiser demonstrates agentic traits. According to the work of Rossi Lamastra and Butticè (2020), who referred to the dictionaries of Pietraszkiewicz and colleagues (2019), an agentic language is determined by a specific set of worlds which includes *"accomplishment", "productivity", "winning"*. The feature "*s_communal*", with respect to the previous introduced, is related to the extent to which the fundraiser demonstrates communal traits. Again, citing the dictionaries of Pietraszkiewicz and colleagues (2019), a communal language includes words like *"solidarity", "mutual", "loving"*. Basically, the difference, between these two last predictors, is that the first agentic has a strong link of masculine characteristics, instead, communal shows an association to feminine ones. The work of Butticè and colleagues, previously mentioned, has been used as a reference also to code the variable "*s_green*". Indeed, in their analysis, the authors relied on a machine learning algorithm to identify whether

[17] https://liwc.wpengine.com/

a campaign could be considered green or not. Finally, they found that words such as *"environmental", "sustainable", "natural"* are crucial to discriminate between green and non-green projects.

Linguistic features	***Description***
s_agentic	Variable representing the extent to which the fundraisers demonstrate agentic traits.
s_communal	Variable representing the extent to which the fundraisers demonstrate communal traits.
s_green	Dummy variable, equal to 1 if the campaign is considered "green"; 0 otherwise.
s_i	Variable representing the presence of the "I" pronoun.
s_money	Variable representing the presence of money-related words.
s_percept	Variable representing the presence of perceptual words.
s_risk	Variable representing the presence of risk-related words.
s_tone	Variable measuring the tone positivity in relation to the words used.
s_wordcount	Variable representing the number of words used within a campaign text.

Table 18: Linguistic features

5.2.2 Numerical static features analysis

After having introduced all the static variables chosen to conduct our research, we performed a statistical analysis, observing their distributions and characteristics. In the following table (19), we show some descriptive statistics of the numeric static variables, realised using the software *"STATA"*[18], are reported.

In order to obtain all the summary statistics needed, the function "*sum, d"* [19]was used on each variable of interest. For each predictor, the function reports the values of the *maximum* and *minimum*, the relative location of the *25th, 50th, 75th percentile*, the values of *mean*, *standard deviation* and those described symmetry, i.e., *skewness* and *kurtosis*.

All the values obtained through the analysis are shown in the table below (19).

Variable	***Max***	***Min***	***Q1 [25th percentile]***	***Q2 [50th percentile]*** ***Median***	***Q3 [75th percentile]***
s_backed	1	0	0	0	1
s_country	1	0	0	1	1
s_created	35	0	1	1	2
s_duration	60	7	30	30	35
s_gender	1	0	0	0	1
s_picture	143	0	1	6	14

[18] Stata: Software for Statistics and Data Science | Stata

[19] https://www.stata.com/statalist/archive/2009-03/msg00861.html

s_serial	1	0	0	0	0
s_target	2129084	42,72	1710,19	5530,50	18584,97
s_team	1	0	0	0	1
s_video	9	0	0	1	1

Variable	***Mean***	***Standard deviation***	***Skewness***	***Kurtosis***
s_backed	0,49	0,50	0,02	1,00
s_country	0,63	0,48	-0,52	1,27
s_created	2,27	3,50	5,42	40,90
s_duration	32,34	11,60	0,74	3,92
s_gender	0,27	0,45	1,02	2,04
s_picture	10,77	14,46	3,47	24,25
s_serial	0,24	0,43	1,19	2,42
s_target	29652,07	146009	11,26	143,70
s_team	0,30	0,46	0,90	1,80
s_video	0,91	0,93	3,39	25,15

Table 19: Descriptive statistic - static variables

Again, the same procedure with the use of *sum, d* in the STATA software has been made for all the linguistic features introduced in the *paragraph 5.2.1.* The table below (20) reports their descriptive values.

Variable	*Max*	*Min*	*Q1 [25th percentile]*	*Q2 [50th percentile]* *Median*	*Q3 [75th percentile]*
s_agentic	28	0	4	7	10
s_communal	25	0	2	5	8
s_green	1	0	0	0	1
s_i	11,54	0	0	0,51	2,7
s_money	9,57	0	0,60	1,08	1,95
s_percept	8,64	0	1,21	1,94	2,83
s_risk	2,8	0	0,18	0,39	0,68
s_tone	99	1	61,54	79,14	92,4
s_wordcount	5257	78	309,5	538	857,5

Variable	***Mean***	***Standard deviation***	***Skewness***	***Kurtosis***
s_agentic	7,17	4,67	0,93	4,43
s_communal	5,79	4,34	1,02	4,30
s_green	0,30	0,46	0,87	1,75
s_i	1,73	2,44	1,76	5,89
s_money	1,40	1,16	2,01	10,73
s_percept	2,22	1,49	1,30	5,10
s_risk	0,48	0,45	1,57	6,45
s_tone	73,22	23,57	-1,09	3,60
s_wordcount	696,87	610,03	3,06	18,29

Table 20: Descriptive statistic – linguistic features

After having comprehended the above values, it is evident that for some features, those data related to symmetry, namely *skewness* [20]and *kurtosis*[21], present some anomalies because they do not fall within their optimal range. For these reasons, we decided to further analyse the predictors presenting some anomalous values in the statistical indicators.

Indeed, skewness refers to a distortion which deflects from the symmetrical curve of the normal bell: if the curve is moved to the left or to the right, it is said to be skewed.

[20] https://www.investopedia.com/terms/s/skewness.asp

[21] https://www.investopedia.com/terms/k/kurtosis.asp

Skewness can be quantified as the extent to which a given distribution varies from a normal one, and the acceptable skewness range is from –2 to +2.

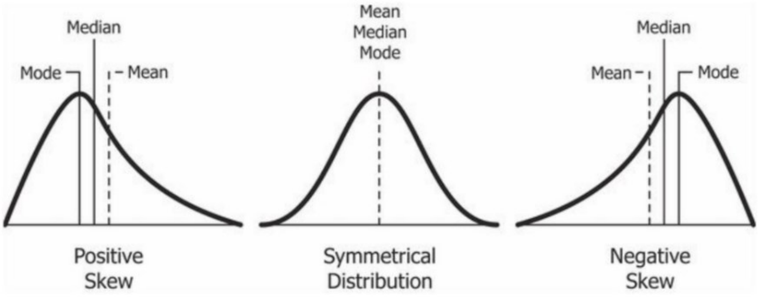


Figure 27: Skewness

Kurtosis is a statistical measure that shows how much both tails of a distribution differ from the ones of a normal distribution. The acceptable range of Kurtosis' values for a variable is between -7 and +7. If a value deviates much from these extremes it is considered anomalous and needs further deepening.

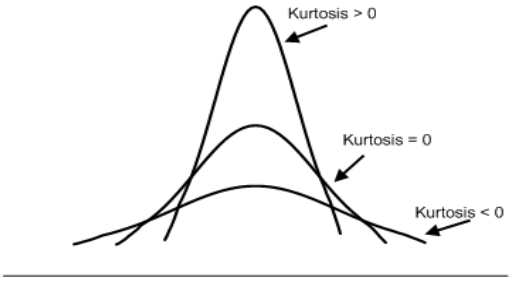


Figure 28: Kurtosis

From the statistical analysis the values that do not lie in the acceptable ranges for both parameters are: *s_created, s_picture, s_target, s_video, s_wordcount* (ANNEX 1 for other variables).

Starting from *s_created*, this variable shows a high asymmetry in the distribution, indeed reports a skewness index of 5,42 which highlights that the distribution is screwed to the right, and a kurtosis index of 40,9, which means that the variable exhibits tailed data exceeding the tails of the normal distribution. The asymmetry of this variable is also confirmed by the fact that the mean is equal to 2,27, showing that on average 2,27 projects have been already created by a fundraiser; but the median has a value equal to 1 and also the third quantile displays a value of 2, both lower than the mean. So, the mean is heavily affected by the presence of some campaigns created by fundraisers that have already previously created a big number of campaigns. Below, the asymmetry of this variable is also corroborated by the boxplot chart reported (30), since the line representing the median is not in the middle of the box. Moreover, the boxplot highlights the presence of several outliers.

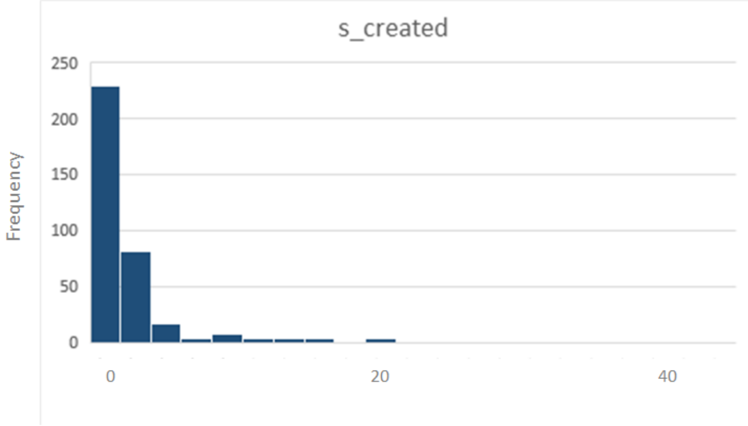


Figure 29: s_created – histogram

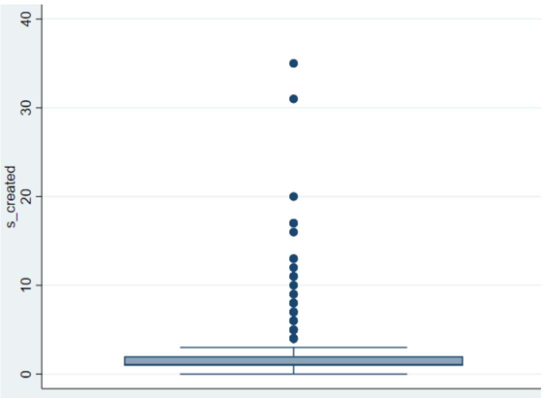


Figure 30: s_created – boxplot

As per the variables *s_picture* and *s_video*, they both have a positive skewness higher than +2 (respectively 3,47 and 3,39), meaning that they have a distribution screwed to the right, as it is also confirmed from the boxplot, since the third quartile is further away from the median compared to the first quartile. Moreover, both the features have a value of kurtosis higher than 20 (respectively 24,25 and 25,15), which suggests that these variables have a *hyper-normal distribution* (i.e., values close to the mean have a higher frequency than the one we would expect in a normal distribution). On average each campaign has nearly 11 pictures, while the number of videos is much lower; indeed, on average, the number of videos is 0.91.

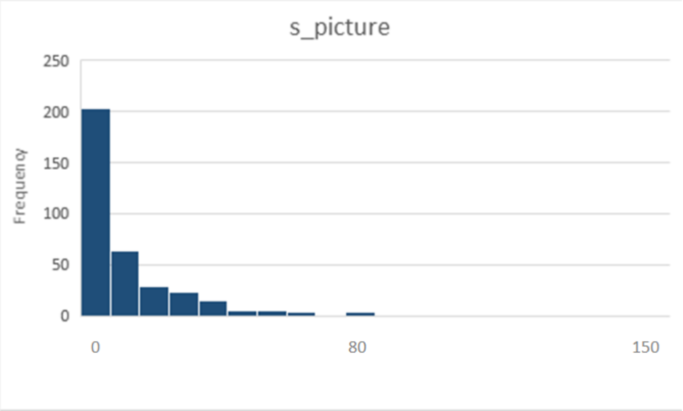


Figure 31: s_picture – histogram

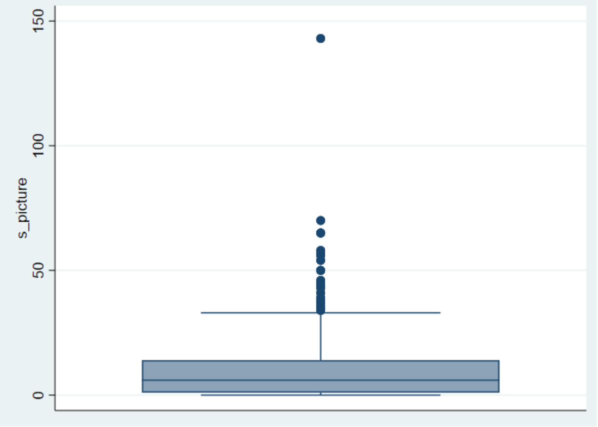


Figure 32: s_picture – boxplot

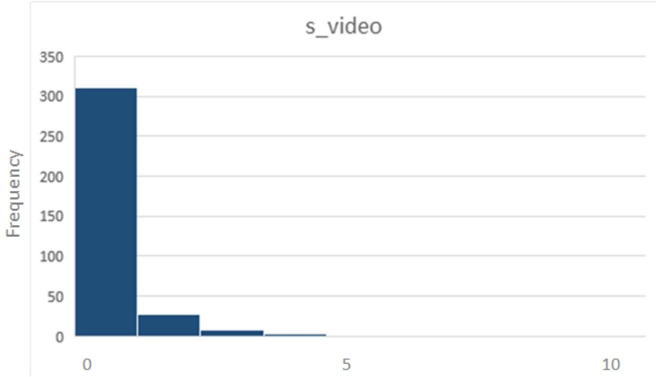


Figure 33: s_video – histogram

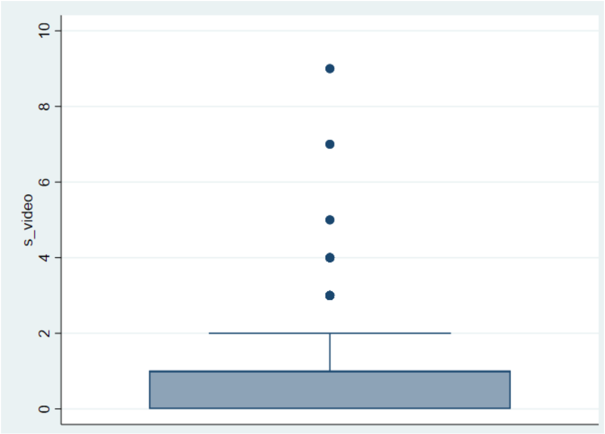


Figure 34: s_video – boxplot

The variable *s_target* shows skewness and kurtosis values highly outside the acceptable ranges. The skewness indicator strongly highlights a distribution skewed to the right, while the value of the kurtosis index underlines a hyper-normal distribution. Moreover, values are highly dispersed, and this is also highlighted by the high standard deviation of this variable.

To better show the distribution, we have created the histogram (figure 35) of this variable, which clearly shows a distribution skewed to right.

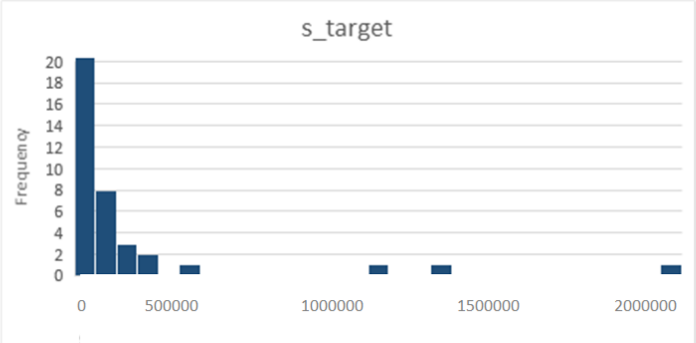


Figure 35: s_target – histogram

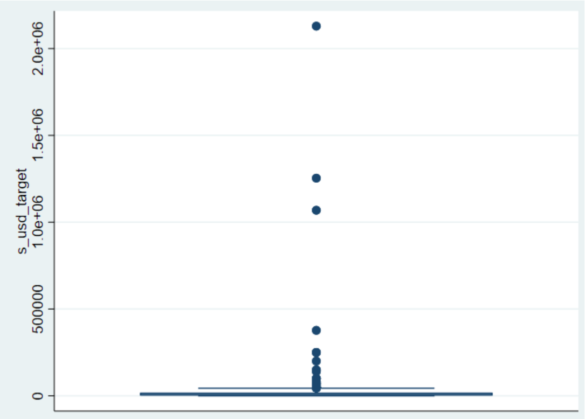


Figure 36: s_target – boxplot

The figure 36 shows the boxplot of the variable *s_target*, where we detected several outliers, as many campaigns present values exceeding the higher whisker of the boxplot.

Among the linguistic variables, we find *s_wordcount* being asymmetrical, indeed it shows a value of skewness equal to 3,06 (namely the distribution is quite skewed to the right, as also confirmed by the histogram (37)) and kurtosis equal to 18,09. Moreover, this variable is highly dispersed, indeed it shows a standard deviation value equal to 610,03. On average each campaign reports 697 words.

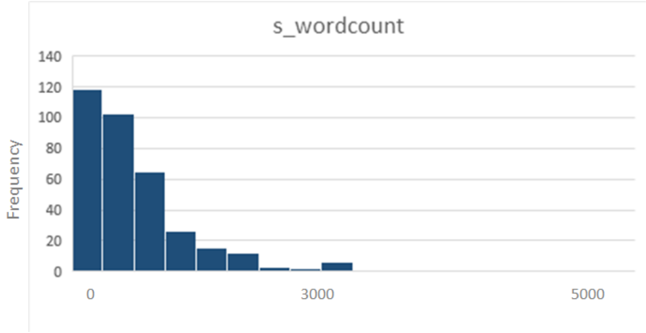


Figure 37: s_wordcount – histogram

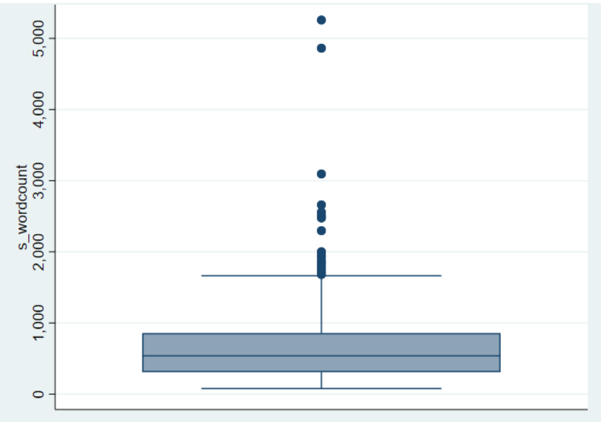


Figure 38: s_wordcount – boxplot

As shown in the boxplot in figure 38, the variable *s_wordcount* shows several outliers.

5.2.3 Dynamic variables

We define as dynamic variables the features that assume different values during the life-cycle of a campaign. As for the static variables, Annunziata and Aversa (2020) collected and stored daily data to deepen the performances that the crowdfunding campaigns reach every day. These data, which are not immediately available on Kickstarter, have been found through the website *Kicktraq*.[22]

[22] Kicktraq

The dynamic variable that we derived from their work are: *d_backers*, *d_percentpledge*, *d_comments* and *d_updates*.

All of them are cumulative variables, which implies that for each day considered, the resulting value is obtained as the sum of all the values of the previous days plus the quantity of the latest day taken into account. Moreover, it is worthy to precise that *d_backers*, *d_comments* and *d_updates* are expressed in absolute terms, instead, *d_percentpledge* is a percentual variable, calculated as pledge reached on required target amount set by the fundraiser. Indeed, in this way, the actual *"success"* or *"over-success"* of the campaigns is reliable, since the campaigns in our sample set very different target goals, some very high and others very low, and without a *"percentage"* value, absolute terms would have led to several bias for subsequent analyses.

Precisely, *d_backers* represents the number of backers contributing to the campaign in the different days; it is computed as follow:

$$\text{d\_backers (t)} = \sum_{k=1}^{t} d\_backers(k)$$

$$\text{with } 1 \leq t \leq T$$

d_percentpledge is the ratio between the amount pledged during the campaign duration and the target (set at the beginning by the fundraiser, static); it is computed as follow:

$$\text{d\_percentpledge (t)} = \frac{\sum_{k=1}^{t} d_{percentpledge}(k)}{s\_target}$$

$$\text{with } 1 \leq t \leq T$$

d_comments identifies the number of comments the campaign receives in the dedicated section during the days; it is computed as follow:

$$\text{d\_comments (t)} = \sum_{k=1}^{t} d\_comments(k)$$

$$\text{with } 1 \leq \text{t} \leq \text{T}$$

d_updates constitutes the number of updates that the fundraiser uploads day by day on the platform; it is computed as follow:

$$\text{d\_updates (t)} = \sum_{k=1}^{t} d\_updates(k)$$

$$\text{with } 1 \leq \text{t} \leq \text{T}$$

In all the different formula, *"T"* represents the whole duration of the campaign, which on Kickstarter can reach a maximum value of 60 days.

Dynamic variables	***Description***
d_backers	Variable representing the sum of the number of backers that up to day t have supported the campaign.
d_comments	Variable representing the sum of the number of comments that the campaign has received up to day t.
d_updates	Variable representing the sum of the number of updates that the fundraiser has uploaded up to day t.

d_percentpledge	Variable representing the sum of the pledged money received by the campaign up to day t over the target goal (*s_target*).

Table 21: Dynamic variables

Chapter 6 – Methodology and empirical findings

The methodology adopted to achieve the objectives set in *Chapter 3*, consists of two main phases. Indeed, in order to build a classification model capable of discriminating overfunding in a very accurate way, a double binary analysis was carried out.

Specifically, the first binary analysis was resumed by the Dissertation *"Predicting the success of crowdfunding campaigns"* developed in 2020 by Annunziata and Aversa. Basically, their studies raised questions about the possibility of predicting the success of a campaign on the basis of the determinants of crowdfunding success, assuming a dynamic perspective. Indeed, along this line of reasoning, the research focused on the possibility of developing a classification tool, able to predict the success or failure of a campaign starting from a set of predictors (Annunziata and Aversa, 2020).

The second binary analysis carried out is at the heart of this research work, since proposes to build a classification model able to discriminate between success and *"over-success"*, in order to answer properly to our research objectives. For the model implementation, all unsuccessful campaigns were excluded from the database, and only the remaining 157 campaigns were used (the database is explained into details in *paragraph 5.2.*). The latter have reached, by the closing date of the project, a pledge equal or greater than the required target. By exploiting them, we developed a classification tool able to predict the success or overfunding of a campaign through a set of specific features.

In the next sections we will report firstly the salient traits of the binary analysis between success and failure of a campaign derived by the work of Annunziata and Aversa (2020) that will be insightful for our analysis, then a detailed description of the steps followed to answer our research questions.

6.1. First binary analysis: Unsuccess vs Success

As we pointed out, our analysis started with previous considerations made by Annunziata and Aversa (2020). Indeed, their work *"Predicting the success of crowdfunding campaigns"* started using a dataset composed by 352 projects, with the willingness to build a classification model able to discriminate between unsuccess (0) and success (1). The practitioners, in order to operationalise the campaign's outcome created the dummy *s_success:* the dependent variable assumes a value of 1 if the campaign meets the target goal before the closing date, 0 otherwise. This response variable was the predicted class label associated to a campaign, which was determined by the classification models according to the predictors given as input.

Dataset and Predictors

The dataset of Annunziata and Aversa (2020) was comprehensive of all the 352 Kickstarter campaigns, and given this sample size and a low risk of overfitting in the construction of the model, they used 24 determinants, all the 20 static variables and all the 4 dynamic predictors already presented (*Paragraphs 5.2.1* and *5.2.2*) Specifically, they decided to consider the first 7 days of the projects to evaluate the dynamic features. Indeed, as mentioned in their thesis, Annunziata and Aversa (2020) focused on the first week of the crowdfunding campaigns, choice driven by the extant crowdfunding studies, which demonstrate the importance of the first seven days of a campaign for determining its success or failure (Petitjean, 2018).

Main insights obtained

The classification model built by the two practitioners could lead to several implications: first of all, as reported in their Dissertation, the work corroborates previous results according to which the first week of a crowdfunding campaign is crucial and it is very informative for predicting the outcome of a campaign (Annunziata and Aversa, 2020). Moreover, when computing variables' importance in foreseeing the success probability, their analysis revealed that when only static predictors are provided to the model, the most important ones are campaigns-related predictors, such as presence of pictures, the target, and the project category. Conversely, linguistic features do not seem as so important. Instead, when the model includes dynamic predictors, the most important one is the percentage of funding target pledged until that moment (Annunziata and Aversa, 2020).

6.2. Second binary analysis: Success vs Overfunding

After deepening the results from the research of Annunziata and Aversa and having verified them properly, we defined the steps to start the construction of a new classification tool, i.e., the one used to discriminate between success and *"over-success"*. Below, all considerations and processes performed.

6.2.1. Overfunding class

As stated in the introduction of *Chapter 6,* we started from a database composed of 352 campaigns, respectively 195 belonging to the *"unsuccess"* category and 157 belonging to the *"success"* one. Among the 157 successful campaigns we made a step forward, trying to discriminate among success and overfunding. For operationalizing the campaign's outcome, since the literature is not clear in providing a comprehensive definition and operationalisation of overfunding, we firstly calculated the overfunding percentage for each campaign, defined as the ratio between pledge and target:

Overfunding percentage: pledge received/ target goal

Then, we defined the dummy variable *overfunding class.* This variable is equal to 1 if the overfunding percentage of the campaign is higher than a defined threshold, 0 otherwise.

Success: if 100%≤ overfunding percentage < Overfunding Threshold = 0

Overfunding: if overfunding percentage ≥ Overfunding Threshold = 1

In order to define this threshold which differentiates among success and overfunding, we considered a range of overfunding thresholds that varied between 120% and 175%, in order to maintain a balance between the two classes in our sample. In fact, using the range of percentages between these two extremes, no case has ever led to two classes more unbalanced than 30-70.

120% ≤ Overfunding Threshold ≤ 175%

For each considered percentage, a certain number of campaigns belonged to one class or the other, and in the following table (*table 22*) we show the number of campaigns pertaining to each class and relative percentage.

Overfunding Threshold	***#Success***	***#Overfunding***	***%Success***	***%Overfunding***
120%	58	99	37%	63%
125%	71	86	45%	55%
130%	76	81	48%	52%
135%	81	76	52%	48%
140%	85	72	54%	46%
145%	87	70	55%	45%

150%	91	66	58%	42%
155%	96	61	61%	39%
160%	97	60	62%	38%
165%	99	58	63%	37%
170%	101	56	64%	36%
175%	105	52	67%	33%

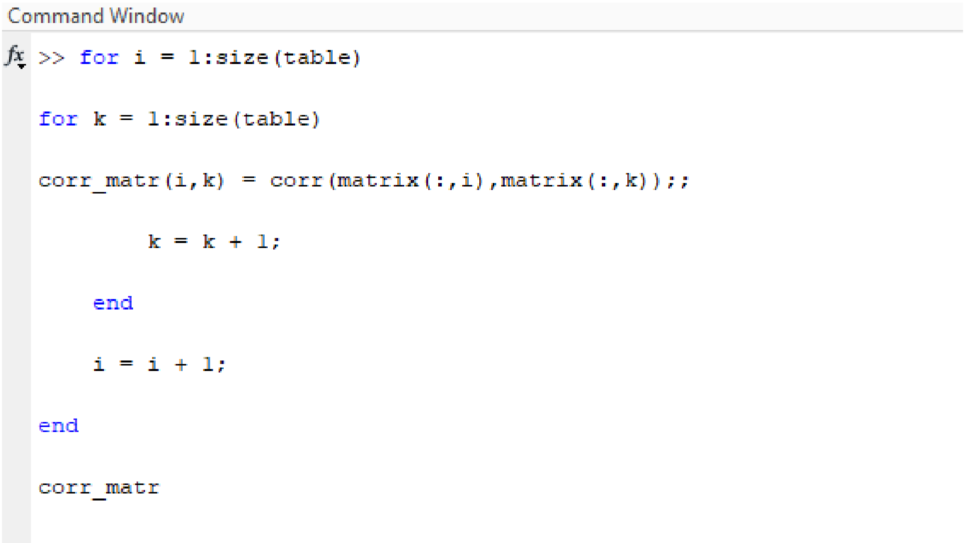
Table 22: Overfunding Class

6.2.2. Correlation analysis

As previously highlighted, the sample used to conduct the second phase of our work, namely the discrimination among success and overfunding, is composed by 157 campaigns. For this reason, since the sample is smaller compared to the initial database of 352 campaigns, we could not use all the 24 static and dynamic variables as predictors to run the analysis.

Indeed, as explained in *Chapter 4* about machine learning methods, practitioners, while developing a classification model, could often run into the problem of *overfitting*. Specifically, this problem arises when the built model learns too well the data used for training, *"learning them by heart"*, resulting in a poorly performance with the testing data, thus losing its generality. The risk of overfitting increases with the number of attributes considered: the higher the number, the more likely it is to run into an irrelevant attribute that soils data with a false pattern. Therefore, the presence of irrelevant attributes could completely hide the relevant ones in the machine learning process.

Following this reasoning, to lower the number of predictors considered, we performed a correlation analysis on the static numerical variables, so as to identify redundant or irrelevant features. In particular, we have set a "*for loop*" [23]on MATLAB, in order to carry out the correlation analysis between all the values of our static variables. Below (figure 39) we show the implemented code, whose output was the *Correlation matrix* (19 x 19).



```
>> for i = 1:size(table)

for k = 1:size(table)

corr_matr(i,k) = corr(matrix(:,i),matrix(:,k));;

        k = k + 1;

    end

    i = i + 1;

end

corr_matr
```

Figure 39: Matlab correlation analysis

In order to have a graphical representation of the Correlation matrix, we employed the MATLAB function "*imagesc (corr_matr)*[24]" which provides a visualization of the correlation between variables (figure 40). The values on the two axes range from 1 to 19, namely the static numeric variables of our dataset. The colours displayed on the correlation matrix show the strength of the relationship between two variables according to the correlation coefficient. The yellow colour on the diagonal indicates a coefficient equal to 1, given by the fact that on the diagonal of the matrix the correlation is computed between a variable and itself.

[23] https://www.mathworks.com/help/matlab/ref/for.html
[24] https://www.mathworks.com/help/matlab/ref/imagesc.html

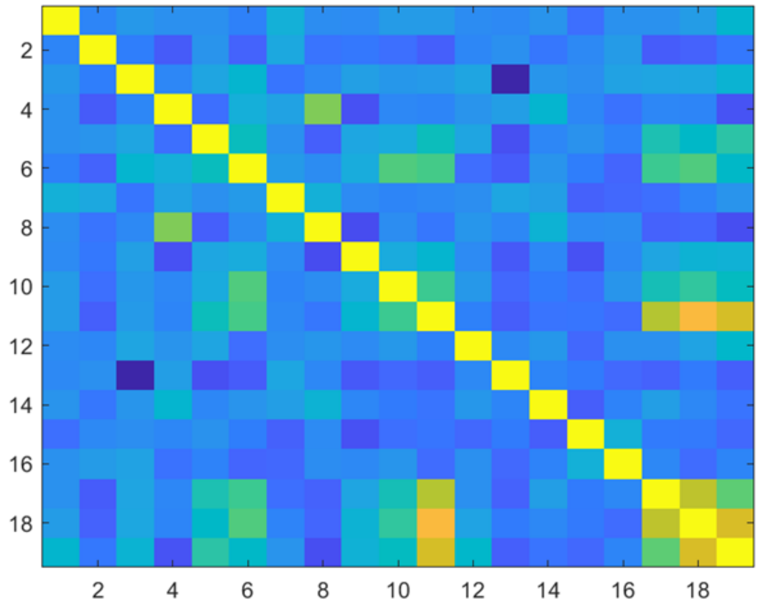


Figure 40: Correlation matrix obtained through the function imagesc(corr_matr)

As easily viewable from the image above, the variables do not have a significant correlation. So, in order to identify which features to eliminate, we set a low threshold equal to 0.4 for evaluating the correlation between two variables, since the risk of overfitting was still too high given the reduced number of campaigns available. In table 23 we provide the pairs of variables with a correlation coefficient higher than 0.4.

Correlated variables	***Correlation coefficient***
s_picture and *s_green*	0.4007
s_picture and *s_wordcount*	0.4235
s_picture and *s_agentic*	0.4411
s_picture and *s_target*	0.4427
s_serial and *s_created*	0.5212
s_wordcount and *s_green*	0.6036

s_wordcount and *s_communal*	0.6687
s_wordcount and *s_agentic*	0.7594

Table 23: Correlation coefficients

At this point we decided to eliminate highly correlated variables, namely:

- *s_created,*
- *s_picture,*
- *s_green,*
- *s_agentic*
- *s_communal*

getting to a total number of variables (also considering the dynamic ones) equal to 19, still too high considered the overfitting risk.

For this reason, we conducted a qualitative analysis and reasoning, in order to identify some features which could be eliminated since they were not fundamental for the scope of our research. As stated in *Chapter 3*, our research goal is to investigate the relative importance of overfunding factors, considering them in an aggregated way to understand their effect on the outcome of a campaign. For this reason, we preferred, as a first contribution to the existent literature, to select determinants that had already been addressed by scholars in a fragmentary and non-integrated way.

For this reason, also supported by the fact that from the previous work of Annunziata and Aversa (2020), it emerged that linguistic variables weigh less than campaign-related and fundraiser-related ones in predicting the outcome of a campaign, we decided to eliminate most of the linguistic features, with the exception of *s_wordcount.* namely:

- *s_I;*
- *s_risk;*
- *s_tone;*
- *s_money;*
- *s_percept;*

So, our analysis will proceed considering a total of 14 determinants.

6.2.3. Time horizon

Kickstarter, reward-based crowdfunding platform, allows fundraisers to publish campaigns with a duration characterised by a minimum of 1 day, up to a maximum of 60 days.

While regarding the success, it is clear in the literature that the first week of the project has a fundamental impact to predict the possible success or failure of a campaign, for the theme of overfunding the *"time horizon"* still remains an obvious gap not already addressed by scholars. For this reason, since we do not have any *"reliable"* information about the crucial days to predict overfunding, we decided to consider the dynamic variables over the entire duration of the campaigns.

Specifically, in our sample of 157 campaigns (the ones used to discriminate between success and overfunding) we realised that the projects collected did not have the same duration. For this reason, we decided to deepen the number of active campaigns as the days progress until the 60 days available. As shown in the figure below (figure 41), the thirty-first day taken into analysis records a drastic decrease of the still active campaigns, precisely only 47 campaigns are still open (30% of the initial sample). This implies that, since the majority of the campaigns close by the end of the 31st day, from that moment on, the sample becomes too small and thus with a low research interest.

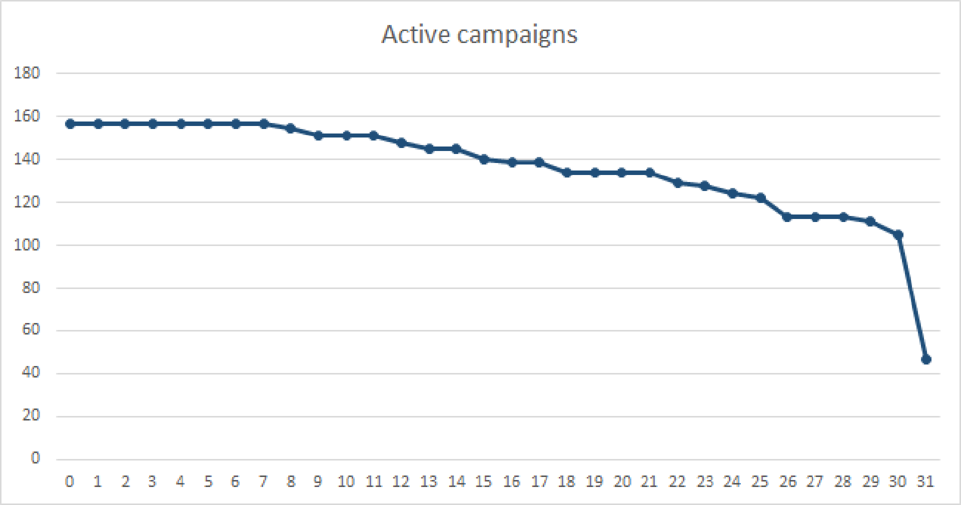


Figure 41: Active campaigns trend during days

Therefore, in order to make our analysis homogeneous and consistent, we thought to consider how the values of the dynamic variables *d_backers*, *d_comments*, *d_percentpledge* and *d_updates* varied until the 31st day. In depth, for each day, it is interesting to see how the average value of the dynamic features begins to decrease after a certain threshold, *"settling"* after the 20th day (*Figure 42, 43, 44, 45*). The trend of the daily variation again justifies the choice not to consider dynamic values after the 31st day, when many closed campaigns no longer have data about their daily number of backers, updates, comments and pledge received.

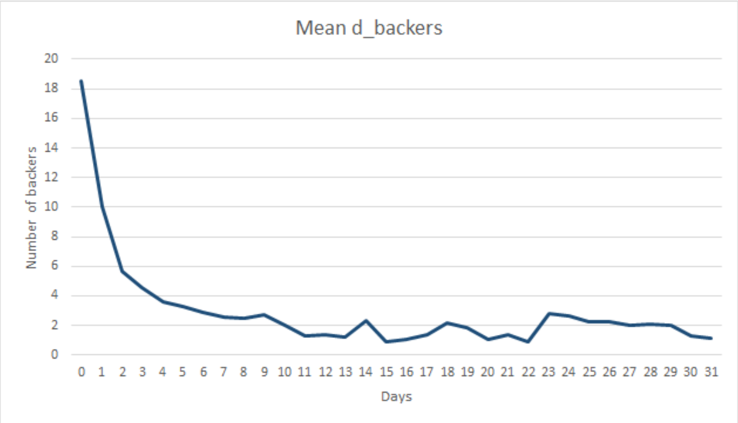


Figure 42: Mean d_backers

Starting with the number of backers, it is possible to notice that the number of people investing on the project, after an initial peak, decreases as the days go by. In particular, after the 23rd day, the data start to decrease reaching very low values.

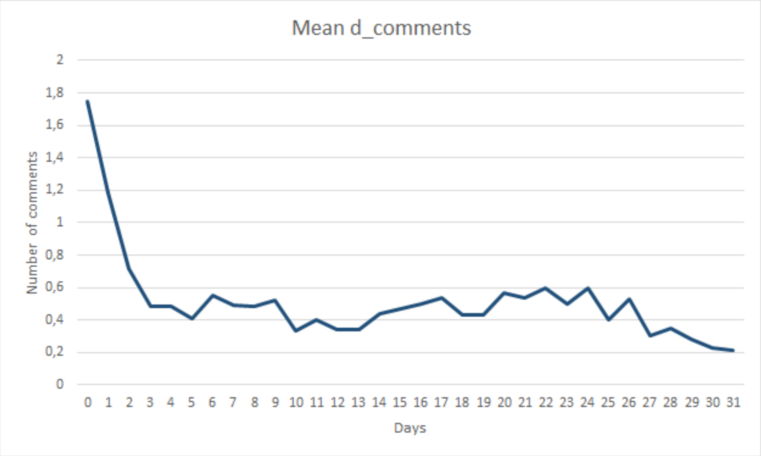


Figure 43: Mean d_comments

A similar reasoning can be made for the number of comments, which have a high peak on the first day and then decrease. In fact, the average daily number of comments from the second day onwards is always less than 1. This is because several projects have never received one during the entire duration of the campaign.

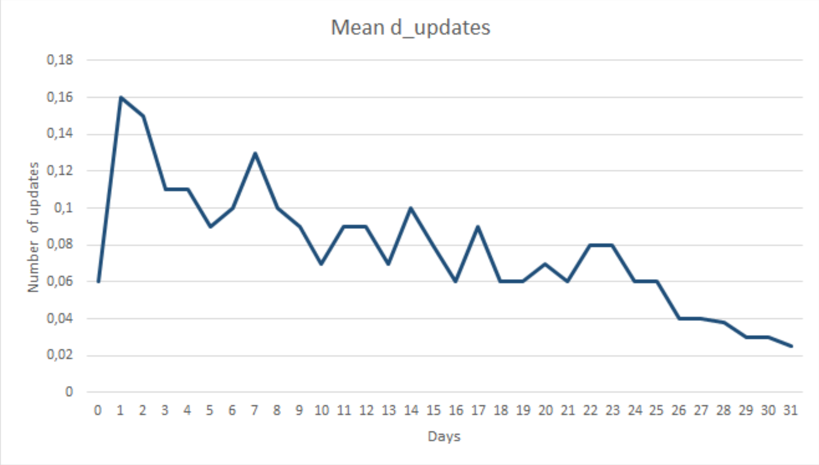


Figure 44: Mean d_updates

As per the number of updates, it can be noticed by the figure above, the mean never exceeds a value of 0,2 for any of the days considered, with a decreasing trend the closer we get to day 31$^{st.}$

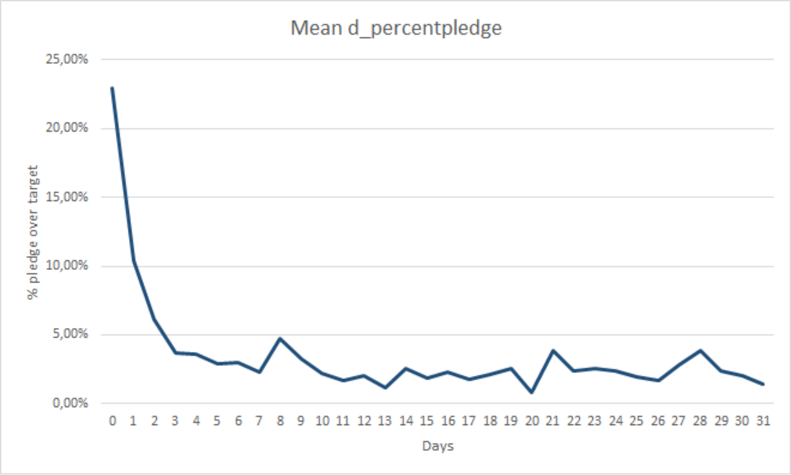


Figure 45: Mean d_percentpledge

Also, for the dynamic variable *d_percentpledge* we can assess a decrescent trend, in particular from day 7 the value of this variable never exceeds 5%.

Moreover, it is worthy to underline that our choice to focus on the values of the dynamic predictors for 31 days does not conflict with the third goal of the research, namely to identify if there are days or weeks crucial to achieve overfunding (*Chapter 3*). Indeed, as reported in the following table (24), almost all the campaigns presented in the dataset that overcome a certain overfunding threshold, among all the variations considered in the range [120%, 175%], reach that percentage within 31 days. Therefore, the values of the table demonstrate that, to evaluate which days are crucial for overfunding, it is enough to focus on the first 31 days.

Overfunding Threshold	***#Overfunding***	***#Overfunding dd 31***	***% dd 31***
120%	99	92	92,93%
125%	86	82	95,35%
130%	81	76	93,83%
135%	76	71	93,42%
140%	72	69	95,83%
145%	70	62	88,57%
150%	66	62	93,94%
155%	61	60	98,36%
160%	60	59	98,33%
165%	58	57	98,28%
170%	56	54	96,43%
175%	52	52	100%

Table 24: Number of campaigns reaching overfunding per Overfunding class

In support of everything we have previously said, in the figures below (46, 47, 48, 49) we report how the number of campaigns that reach overfunding varies in the different days considered, for four overfunding percentages, respectively 135%, 145%, 165% and 175% (ANNEX 2 for other percentages). As it can be seen from the graphs, the number of overfunding campaigns reaches asymptotically 100% on the 31st day, in support of the fact that for the phenomenon of overfunding it is sufficient to consider, indeed, 31 days.

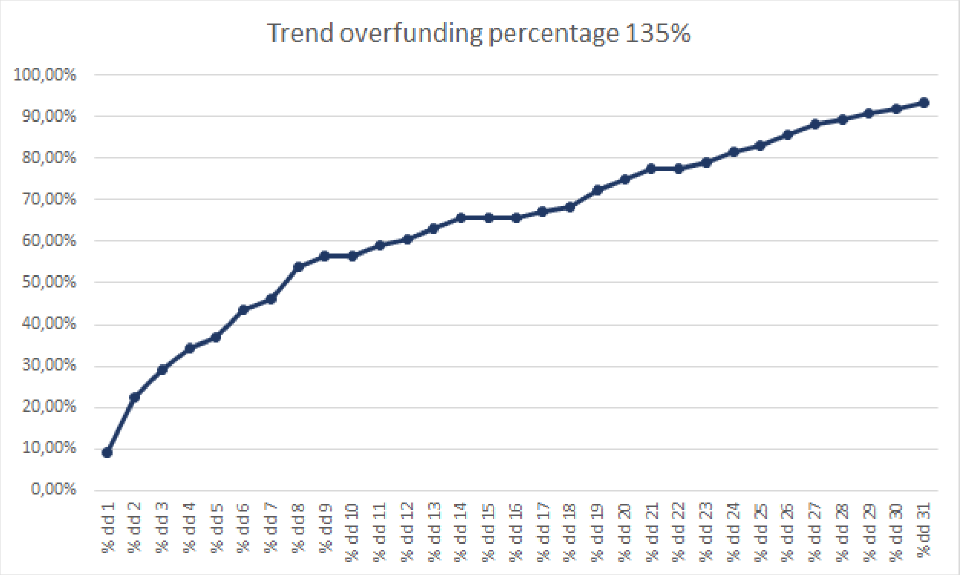


Figure 46: Trend overfunding percentage 135%

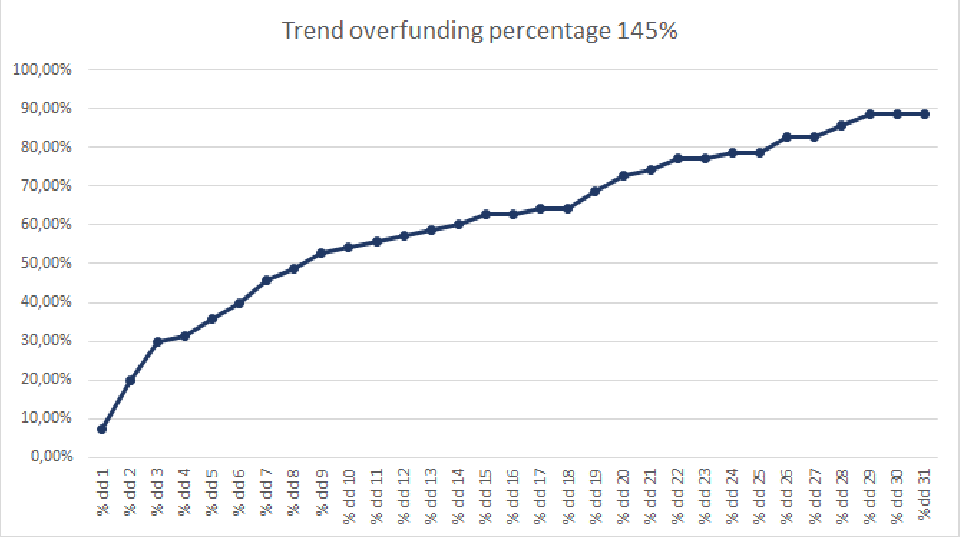


Figure 47: Trend overfunding percentage 145%

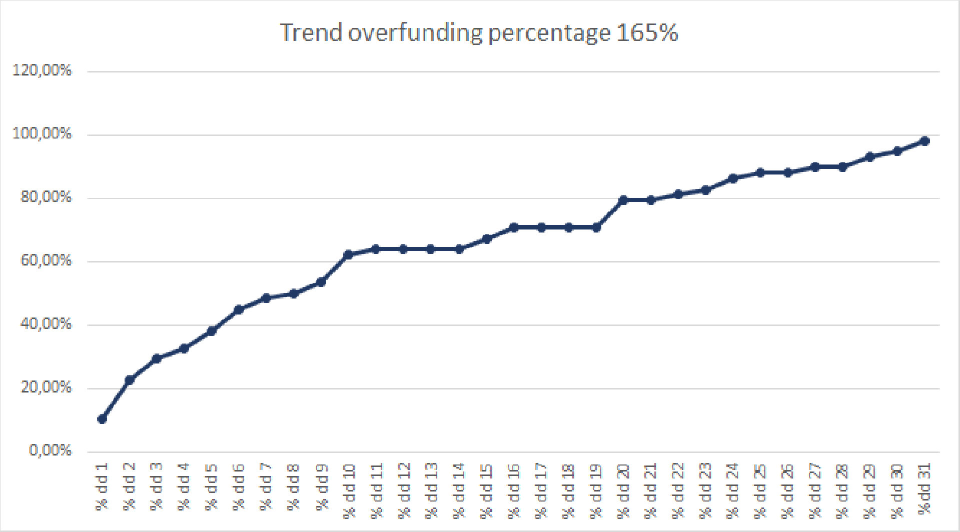


Figure 48: Trend overfunding percentage 165%

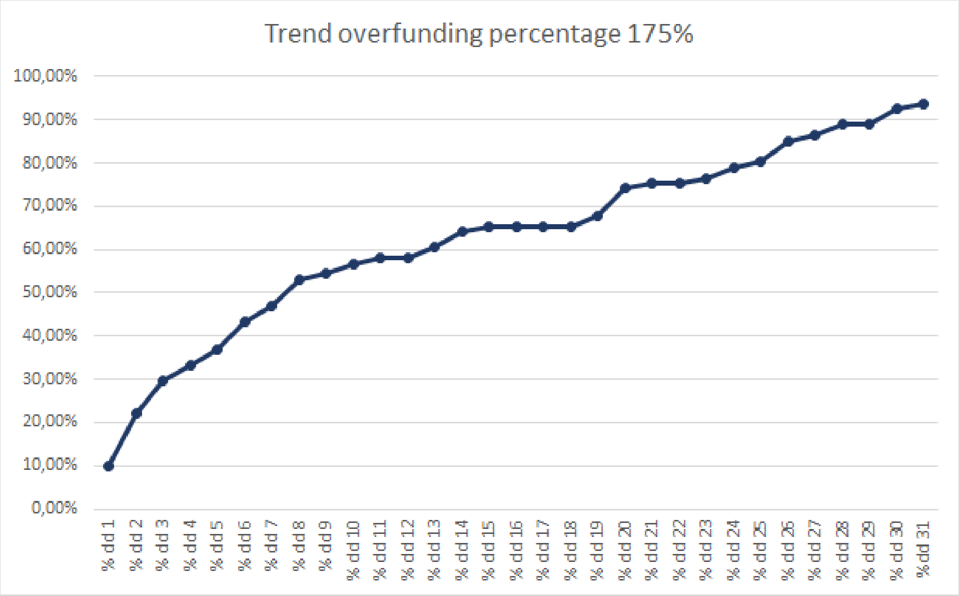


Figure 49: Trend overfunding percentage 175%

6.3. Development of a Classification Model

6.3.1. Data preparation: Data validation and Transformation

After all the premises made in the previous paragraphs, it is possible to define the final dataset used to set the classification model. To recap, we selected 10 static variables and 4 dynamics after the correlation analysis, and defined a time range of 31 days. Thus, we built 32 datasets, for each of the 12 overfunding percentages considered, for a total of 384 datasets. Precisely:

- 1 *static dataset* including static variables (which represents the values of the launch day, prior to the opening of the campaign), The static dataset has, for each 157 campaigns (*lines*), 10 static variables available at the beginning of the campaigns (*columns*);
- *31 dynamic datasets*, each including the dynamic variables associated with each day of the campaign. To all the 157 observations (*lines*) 14 predictors are associated, precisely 10 static and 4 dynamics variable (*columns*).

The choice of building 32 different datasets is driven by the fact that, in order to build a classification tool exploiting machine learning algorithms, day by day a new set of dynamic information should be provided to the learning technique. On a daily basis, the predicting models will elaborate the new dataset, which is cumulative and considers the values up to the day taken into account.

Before starting to import datasets in the MATLAB software, a procedure of data preparation and pre-processing has been developed, following the steps and rules already presented in *Chapter 4* (*"How to build a classification model"*).

Indeed, many machine learning algorithms propose to find trends in data by comparing values of the data points. But typically, in the majority of the cases, variables are measured at different scales (i.e., dummies, categorical, numerical) and thus, they do not equally weight to the model fitting. For this reason, to avoid that the model ends up with bias, before providing the data to the model, a normalization is needed. The most used standardisation techniques are *Z-Score Normalisation* and *Min-max*. Among them, we decided to opt for the second one, since with the former the features would not have been on the *exact* same scale.

Having several dummy variables in our dataset of predictors, we wanted all the values of our features ranging from 0 to 1: given that the dummies present already values ranging from a minimum of 0 to a maximum of 1, the normalization of these variables has not been necessary. Below, the formulation of the standardization employed for each value of our datasets:

$$X'_{ij} = \frac{Xij - Xminj}{Xmaxj - Xminj}$$

Where X_{ij} is the starting value to be transformed, $Xmin_j$ is the minimum value of the predictor j and $Xmax_j$ is the maximum value of the predictor j.

6.3.2. Training phase

Before reaching the heart of the training phase, it is necessary to introduce the MATLAB app used to carry out this process, to highlight its main functionalities. Then, all the steps followed to execute the training of the classification tool are presented.

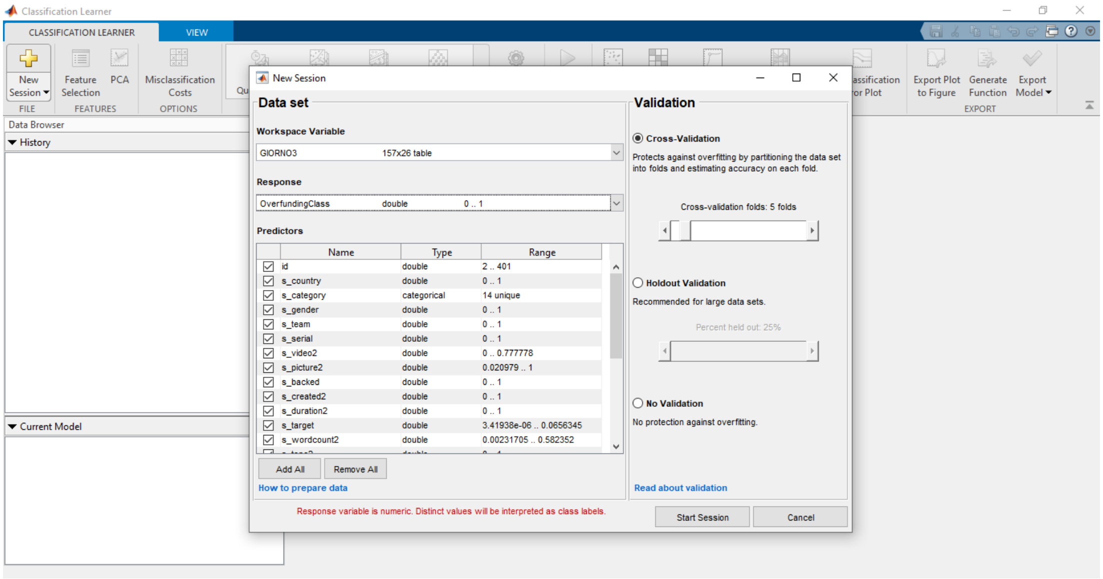
6.3.2.1. Classification learner app overview

As reported in Matlab Help Center[25], *Classification Learner app* lets practitioners perform common supervised learning tasks such as interactively exploring data, selecting features, specifying validation schemes, training models, and assessing results. Specifically, the app chooses among various algorithms to train and validate

[25] https://it.mathworks.com/videos/classify-data-using-the-classification-learner-app-106171.html?requestedDomain

classification models for binary or multiclass problems. After training multiple models, it is necessary to compare their validation errors side-by-side, and then choose the best model.

The *Classification Leaner App* takes as input the dataset previously prepared (*Dataset and Data preparation*), which is the set whose rows are the instances used for training the model (the crowdfunding campaigns) and the columns are the variables. After having provided the dataset as input, a dialog box appears in the app (Figure 50), in which the user has to select the response variable. This, in our particular case, is the dependent variable "*OverfundingClass*". In the same dialog box, the practitioner has to select the *predictors*, namely the features associated to each campaign. By default, the app selects all the available predictors, and it is up to the user to deselect those not to be included in the analysis. Lastly, before starting to train the model, the validation method has to be chosen. As we can also see in picture (50) two validation methods are proposed: *k-fold cross-validation* and *holdout validation*.



New Session

Data set

Workspace Variable

GIORNO3 157x26 table

Response

OverfundingClass double 0 .. 1

Predictors

	Name	Type	Range
☑	id	double	2 .. 401
☑	s_country	double	0 .. 1
☑	s_category	categorical	14 unique
☑	s_gender	double	0 .. 1
☑	s_team	double	0 .. 1
☑	s_serial	double	0 .. 1
☑	s_video2	double	0 .. 0.777778
☑	s_picture2	double	0.020979 .. 1
☑	s_backed	double	0 .. 1
☑	s_created2	double	0 .. 1
☑	s_duration2	double	0 .. 1
☑	s_target	double	3.41938e-06 .. 0.0656345
☑	s_wordcount2	double	0.00231705 .. 0.582352

Add All | Remove All

How to prepare data

Response variable is numeric. Distinct values will be interpreted as class labels.

Validation

● Cross-Validation

Protects against overfitting by partitioning the data set into folds and estimating accuracy on each fold.

Cross-validation folds: 5 folds

○ Holdout Validation

Recommended for large data sets.

Percent held out: 25%

○ No Validation

No protection against overfitting.

Read about validation

Start Session | Cancel

Figure 50: Classification Lerner App dialog box

For the cross-validation method, *k* is set at a value of 5, but it can be changed manually by the user and as reported by MATLAB, this method *"protects against overfitting by* *partitioning the data set into folds and estimating accuracy on each fold"*. On the other hand, MATLAB recommends the holdout validation in case of large data sets.

Once all settings are selected, the *training phase* begins. In the C*lassification Learner app*, different classifiers belonging to several categories can be trained, in order to understand which one works better (in term of accuracy) given the response variable and the predictors provided. For the scope of our analysis, we trained 15 models, presented in the table below (25). The entire environment of the *Classification Learner app* includes other *"model types"* in addition to those reported below, but these models require all the predictors to be a numeric variable, and since we did not exclude through the correlation analysis our categorical feature "*s_category*", the only ones useful to conduct the analysis were the 15 presented in the table.

Classification Category	***Model Type***
Decision Trees	Fine Tree Medium Tree Coarse Tree
Logistic regression classifier	Logistic regression
Naive Bayes Classifiers	Naive Bayes (Gaussian) Naive Bayes (Kernel)
Support Vector Machines	SVM (linear) SVM (quadratic) SVM (Cubic) SVM (Fine Gaussian)

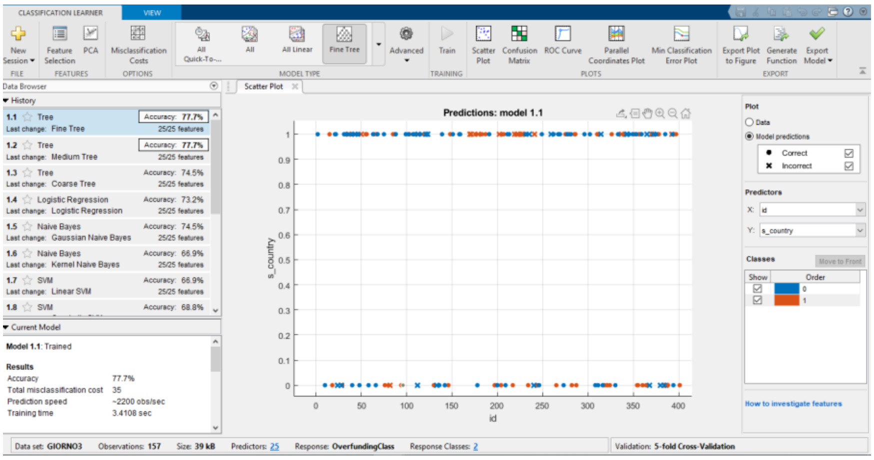
	SVM (Medium Gaussian)
	SVM (Coarse Gaussian)
Ensemble Classifiers	Ensemble (Boosted Trees)
	Ensemble (Bagged Trees)
	Ensemble (RUSBoosted Trees)

Table 25: MATLAB classifiers

As shown in the picture below (51), in the main page of the app, different functionalities are displayed. One of them is the *features selection* where the user can deselect some of the predictors previously included, and see how the accuracy of the model changes accordingly. Then we find the function *Advance* that allows to tune advanced model options.

Moreover, on the app it is possible to visualize the *Confusion matrix, ROC Curve* and *Parallel Coordinates Plot* for each one of the trained models.

Lastly, we find the *Export model* functionality, used to export the trained model which can be then used for the testing phase on a new dataset.



History

1.1 ☆ Tree Accuracy: 77.7%
Last change: Fine Tree 25/25 features

1.2 ☆ Tree Accuracy: 77.7%
Last change: Medium Tree 25/25 features

1.3 ☆ Tree Accuracy: 74.5%
Last change: Coarse Tree 25/25 features

1.4 ☆ Logistic Regression Accuracy: 73.2%
Last change: Logistic Regression 25/25 features

1.5 ☆ Naive Bayes Accuracy: 74.5%
Last change: Gaussian Naive Bayes 25/25 features

1.6 ☆ Naive Bayes Accuracy: 66.9%
Last change: Kernel Naive Bayes 25/25 features

1.7 ☆ SVM Accuracy: 66.9%
Last change: Linear SVM 25/25 features

1.8 ☆ SVM Accuracy: 68.8%

Current Model

Model 1.1: Trained

Results

Accuracy	77.7%
Total misclassification cost	35
Prediction speed	~2200 obs/sec
Training time	3.4108 sec

○ Data
● Model predictions
● Correct ☑
✖ Incorrect ☑

X: id
Y: s_country

Show	Order
☑	🟦 0
☑	🟧 1

Data set: GIORNO3 Observations: 157 Size: 39 kB Predictors: 25 Response: OverfundingClass Response Classes: 2 Validation: 5-fold Cross-Validation

Figure 51: Classification Lerner App main page

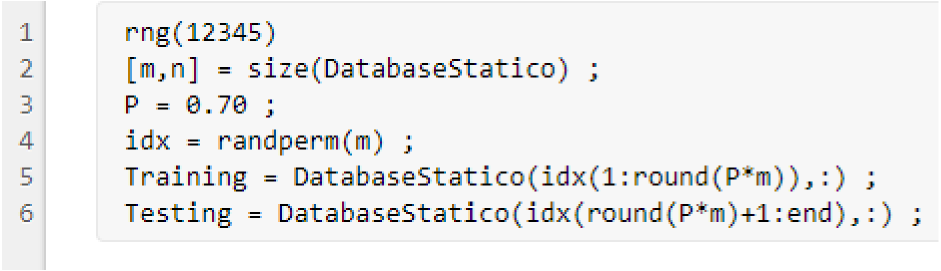
6.3.2.2. Preliminary training-phase

As stated in the *paragraph 6.2.1.,* we defined as output variable of our model, the dummy variable *"overfunding class",* considering different thresholds varying between 120% and 175% in order to understand which of these percentages is more representative of the phenomenon, so as to answer our first research question, namely provide a definition of overfunding (*Chapter 3*).

In order to understand which percentage led to the best accuracy for our model, we performed the training phase through the *Classification Learner app* for each of the *384 datasets* previously created

*384 Datasets = (1 static + 31 dynamics) * 12 (thresholds among 120%- 175% with a difference of 5% between each of them)*

We started the training phase by importing the datasets: we divided the total instances (157 campaigns) in two parts randomly chosen by a MATLAB code, to avoid bias in ours results. Indeed, we decided to use the 70% of the total projects to implement the training of the model (110 campaigns), the remaining 30% to perform the subsequent testing phase at the end. Below, the lines of code implemented on MATLAB to execute these steps are presented:



```
rng(12345)
[m,n] = size(DatabaseStatico) ;
P = 0.70 ;
idx = randperm(m) ;
Training = DatabaseStatico(idx(1:round(P*m)),:) ;
Testing = DatabaseStatico(idx(round(P*m)+1:end),:) ;
```

As it can be easily seen by the MATLAB code reported, the command *rng (12345)* [26]was launched in the *Command Window*. We repeated this setting before starting

[26] https://www.mathworks.com/help/matlab/ref/rng.html

every session on the *Classification Learner app* in order to guarantee reproducibility for each running.

After randomly choosing the 110 campaigns for the testing phase, we selected the 10 features previously defined as predictors (*paragraph 6.2.2. Correlation analysis*). Given the limited size of our dataset, as also suggested by MATLAB itself, we employed as validation method the 5-fold cross validation, so that 1/5 (1/k) of the training dataset is used as fold validation and the app cyclically trains the model, for each validation fold, using the training folds.

In order to assess the best overfunding threshold in terms of accuracy, we trained all the 15 available classification algorithms, recording for each dataset of each percentage the highest accuracy. Indeed, the app provides in a dedicated section the *accuracy*, *total misclassification cost, training time* for each trained model. We performed this procedure for the 1 static dataset and the 31 dynamic datasets, again registering the best accuracy of each percentage. We deeply evaluated the best accuracy of the thresholds and, as reported below (Figure 52), we mapped which was the overfunding threshold with the highest accuracy for the all the 32 datasets. As reported in the table below, the best percentage according to which our model is more performing (in terms of accuracy) is an overfunding threshold of 175%.

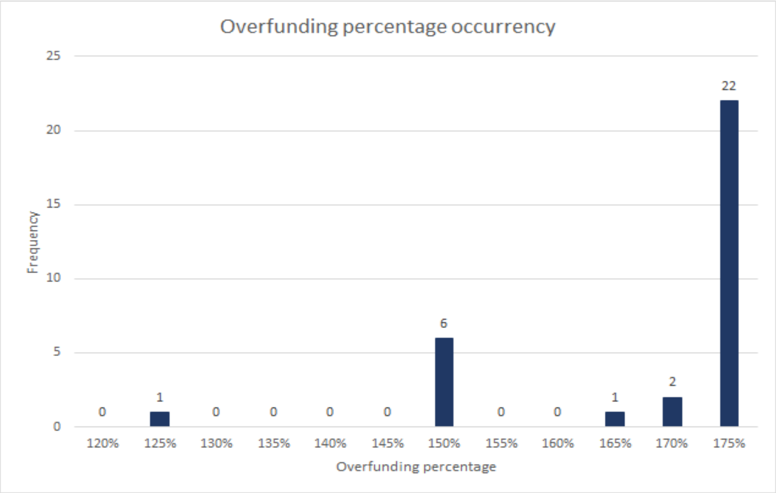


Figure 52: Overfunding thresholds occurrency

6.3.2.3. Training phase execution with an overfunding-threshold of 175%

We proceeded our analysis using the overfunding threshold 175%: we focused our attention on the level of accuracy, and for each of the 32 total datasets, we identified the three most accurate classification models. The Tables (26, 27, 28, 29) provide the results of this phase.

The first column "*Dataset*" refers to the dataset trained (*s* is the static dataset; *d* is the dynamic dataset with t ranging from 1 to 31). The second one, "*Best classifier*" reports the three most accurate classifiers for each dataset, from the first most accurate to the third one; the column "*Model type*" refers to the name of the classifier, and the "*Accuracy*" column displays the accuracy of the classifier.

Dataset	***Best Classifier***	***Model type***	***Accuracy [%]***
s	1	Ensemble RUSBoosted	73,60%
	2	Coarse Tree	70,90%
	3	Ensemble Bugged Tree	67,30%
d1	1	Ensemble Bugged Tree	83,60%
	2	Coarse Tree	80,90%
	3	Naive Bayes Gaussian	80,90%
d2	1	Ensemble RUSBoosted	82,70%
	2	SVM Linear	82,70%
	3	Coarse Tree	81,80%
d3	1	Ensemble RUSBoosted	86,40%
	2	Ensemble Bugged Tree	85,50%
	3	SVM Linear	85,50%
d4	1	Ensemble Bugged Tree	87,30%
	2	Ensemble RUSBoosted	87,30%
	3	Coarse Tree	85,50%

d5	1	Ensemble Bugged Tree	85,50%
	2	SVM Quadratic	84,50%
	3	Naive Bayes Gaussian	83,60%
d6	1	Ensemble Bugged Tree	83,60%
	2	SVM Linear	82,70%
	3	SVM Quadratic	82,70%
d7	1	SVM Quadratic	84,50%
	2	Ensemble RUSBoosted	82,70%
	3	SVM Linear	82,70%

Table 26: Accuracy of the three best classifiers for each dataset (from static database to d7)

Dataset	***Best Classifier***	***Model type***	***Accuracy [%]***
d8	1	Ensemble Bugged Tree	87,30%
	2	Coarse Tree	83,60%
	3	SVM Linear	83,60%
d9	1	Ensemble RUSBoosted	87,30%
	2	Ensemble Bugged Tree	86,40%
	3	SVM Quadratic	83,60%
d10	1	Coarse Tree	87,30%
	2	Ensemble RUSBoosted	85,50%
	3	Fine Tree	85,50%
d11	1	Ensemble RUSBoosted	87,30%
	2	Ensemble Bugged Tree	87,30%

	3	Coarse Tree	86,40%
d12	1	Ensemble RUSBoosted	87,30%
	2	Ensemble Bugged Tree	86,40%
	3	Coarse Tree	85,50%
d13	1	Ensemble RUSBoosted	86,40%
	2	SVM Quadratic	86,40%
	3	Ensemble Bugged Tree	85,50%
d14	1	Ensemble RUSBoosted	87,30%
	2	Ensemble Bugged Tree	87,30%
	3	SVM Quadratic	85,50%
d15	1	Ensemble RUSBoosted	87,30%
	2	Ensemble Bugged Tree	85,50%
	3	Coarse Tree	85,50%

Table 27: Accuracy of the three best classifiers for each dataset (d8 to d15)

Dataset	***Best Classifier***	***Model type***	***Accuracy [%]***
d16	1	Ensemble RUSBoosted	88,20%
	2	Ensemble Bugged Tree	88,20%
	3	Coarse Tree	85,50%
d17	1	Coarse Tree	90,90%

	2	Fine Tree	89,10%
	3	Medium Tree	89,10%
d18	1	Coarse Tree	90,00%
	2	Medium Tree	88,20%
	3	Ensemble Bugged Tree	88,20%
d19	1	Coarse Tree	90,00%
	2	Ensemble Bugged Tree	89,10%
	3	Medium Tree	88,20%
d20	1	Coarse Tree	91,80%
	2	Ensemble RUSBoosted	91,80%
	3	Medium Tree	91,80%
d21	1	Ensemble RUSBoosted	91,80%
	2	Ensemble Bugged Tree	91,80%
	3	Coarse Tree	90,00%
d22	1	Coarse Tree	92,70%
	2	Medium Tree	92,70%
	3	Fine Tree	92,70%
d23	1	Coarse Tree	93,60%
	2	Medium Tree	93,60%
	3	Fine Tree	93,60%

Table 28: Accuracy of the three best classifiers for each dataset (d16 to d7)

Dataset	*Best Classifier*	*Model type*	*Accuracy [%]*
d24	1	Ensemble RUSBoosted	92,70%
	2	Ensemble Bugged Tree	92,70%
	3	Coarse Tree	90,00%
d25	1	Coarse Tree	92,70%
	2	Ensemble RUSBoosted	92,70%
	3	Medium Tree	92,70%
d26	1	Ensemble Bugged Tree	93,60%
	2	Coarse Tree	92,70%
	3	Medium Tree	92,70%
d27	1	Ensemble Bugged Tree	94,50%
	2	Coarse Tree	94,50%
	3	Medium Tree	94,50%
d28	1	Ensemble Bugged Tree	94,50%
	2	Coarse Tree	94,50%
	3	Medium Tree	94,50%
d29	1	Ensemble RUSBoosted	95,50%
	2	Ensemble Bugged Tree	94,50%
	3	Coarse Tree	92,70%

d30	1	Coarse Tree	97,30%
	2	Medium Tree	97,30%
	3	Fine Tree	97,30%
d31	1	Coarse Tree	98,20%
	2	Medium Tree	98,20%
	3	Fine Tree	98,20%

Table 29: Accuracy of the three best classifiers for each dataset (from d24 to d31)

It is difficult to assess at a first glance the most accurate classifiers for the whole 32 datasets. For this reason, to be more consistent, we created another table (30), displaying the number of times each classifier was the first, the second and the third best one, in terms of accuracy. By doing so, we realised that the three most accurate classifiers based on our predictors and response variables are decision tree, specifically *Coarse Tree,* and two ensemble trees, namely *Bagged Trees* and *RUSBoosted Trees*.

	Best Classifier	***Second Best Classifier***	***Third Best Classifier***
Ensemble RUSBoosted	13	5	0
Coarse Tree	10	6	9
Ensemble Bugged Tree	8	11	3
SVM Quadratic	1	2	3
Medium Tree	0	5	7
SVM Linear	0	2	3
Fine Tree	0	1	5
Naive Bayes Gaussian	0	0	2

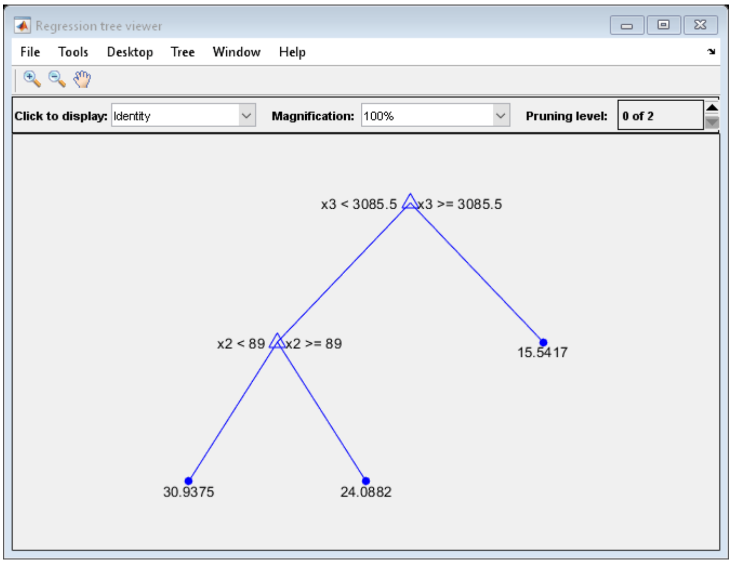
Table 30: Number of times each classification model has been the first, second and third best classifier

After having identified the three most accurate classifiers, we exported our 96 trained models (namely 32 datasets for each best predictors: *Coarse Tree, Ensemble Bagged Tree and RUSBoosted Tree*).

Main characteristics of the best classification models

Coarse tree

The Coarse Tree belongs to the decision trees classifier family, together with Fine Tree and Medium tree. This model is characterised by high prediction speed, namely it is able to foresee the class associated to instances in a very fast way. It is very easy to be interpreted, indeed only few leaves are needed to make coarse distinctions between classes. In particular, to read a response the path from the root node down to a leaf node has to be followed, paying attention to proceed with the right branch at each split. The final leaf node reached represents the response of the prediction, namely the class to be given to the instance. Moreover, differently from both the Fine and Medium tree, the Coarse tree can achieve a maximum of 4 splits, having a very limited depth. In the figure below (53) a graphical example of the functioning of the Coarse tree.



Click to display: Identity
Magnification: 100%
Pruning level: 0 of 2

Figure 53: Coarse tree visual example

Ensemble Bagged tree

Bootstrap aggregation (bagging) is a type of ensemble learning: first, Bootstrap sampling is used to create a subset of data from the training ones, by randomly choosing products with substitution. In this way, it ensures the effect of randomness and independence, which contributed to the model: randomness allows algorithms to avoid overfitting and to better generalize on new observations; moreover, bootstrap effectively helps combat high variance, which is a disadvantage for algorithms such as the decision tree. An example of algorithm using the Bagging technique is the well-known *Random Forest*, that uses the Decision Tree as a basic classifier. Specifically, the algorithm used by Bagged Trees trained through MATLAB is Breiman's *"random forest"*. Random Forest trains many trees on a larger subset of data through bootstrap, in order to have more than one division point while developing the tree, so more perspectives to evaluate and find a model that has better generalization power (more representative of real-world data). Bagging can be computationally expensive, since it requires a longer time to be trained and tend to use more memory space than other classifiers, if the number of estimators is exceptionally high. The interpretability could be negatively affected too, since could be very difficult derive some predictions for each class. Anyway, *Ensemble Bagged* seems to be a fair sacrifice to create a better machine learning model.

Ensemble RUSBoosted tree

RUSBoosted tree belongs to the Ensemble tree family, as Bagged tree. As clearly explained in the MathWorks community, *"RUSBoost is a boosting-based sampling algorithm that handles class imbalance in class labelled data."* In order to achieve this result, the algorithm uses a series of RUS (i.e., random under-sampling) and a very common boosting procedure called *"AdaBoost",* with the aim of better modelling the minority class by removing majority class samples. This algorithm is faster in predicting the classes and also easier to be interpreted compared to the *Bagged tree* classifier. As suggested by the Matlab HelpCenter [27]itself, the *RUSBoosted* classifier is *"good for skewed data (with many more observations of 1 class)"*.

6.3.3. Features' relative importance

In the last paragraph, we discussed how we found out the three best classifiers for the 32 datasets related to an overfunding threshold set at 175%, through the *Classification Learner App.* Now, we proceed to explain how we performed the analysis of the relative importance of the predictors considered, assessing which are the static and dynamic factors that could predict overfunding and which are the most relevant ones (*O2, Chapter 3*).

In the section "*Steps for developing a classification model*", we anticipated that different techniques could be used to identify which are the most relevant predictors: in the end, we developed a *Filter* and *Wrapped* feature selections. In addition, we also provide results about how the accuracy levels change during the training phase (i.e., the accuracy relative to the 5-fold cross-validation on the models), if we train the classification model excluding one given static predictor at a time. This process, as we will see in the dedicated section, is able to demonstrate the importance of each predictor on the accuracy level reached by the model.

6.3.3.1. Filter feature selection

Filter methods are generally used as a pre-processing step, since they work by selecting the features, after having analysed the relationship between the predictors and the response variable, independently from the machine learning algorithm and its biases do not interact with the filter.

We used the selected method on the static dataset and on all the 31 dynamic datasets, to understand how the relative importance of the features changes with the progression of days. To implement this methodology, we used the *Statistics and Machine Learning*

[27] Choose Classifier Options - MATLAB & Simulink - MathWorks Italia

Toolbox of MATLAB, which offers several functions for features selection. In particular, for our problem and data types of our predictors, the most suitable function was *fscmrmr.* Indeed, among all the available ones (i.e., *fscnca, fsrftest, fscmrmr* and others), the one implemented could be adopted to work both with categorical and continuous features. As reported in the Matlab Help Center, the function *fscmrmr*[28] computes the importance of the features for classification using minimum redundancy maximum relevance (*MRMR*) algorithm.

The function receives as input the predictors in the dataset imported and the response variable, and it provides as output an array with a score associated to each predictor. The lines of code implemented in order to perform the analysis using the aforementioned function are the following:

```
idx = fscmrmr (Dataset, "OverfundingClass")'
[idx, scores] = fscmrmr (Dataset, “OverfundingClass”)
```

The first function provides an array with the predictors in descending order of importance, and the second one shows the relative scores. A high score value indicates that the corresponding predictor is important. The following tables (31 to 41) show the score associated to each factor, computed using the function *fscmrmr* for each dataset. In light blue, the dynamic predictors are highlighted.

s		**d1**		**d2**	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
s_serial	0,0416	d_percentpledge	0,2303	d_percentpledge	0,262
s_team	0,0308	s_category	0,0396	s_category	0,0396
s_target	0,0029	s_target	0,0383	s_target	0,0383
s_category	0,0022	s_serial	0,0137	s_serial	0,014
s_gender	0,002	s_team	0,0096	s_team	0,0093

[28] Rank features for classification using minimum redundancy maximum relevance (MRMR) algorithm - MATLAB fscmrmr - MathWorks Italia

s_duration	0,0005	d_comments	0,004	d_comments	0,0044
s_backed	0,0003	s_duration	0,0029	d_updates	0,0039
s_video	0,0002	d_backers	0,0024	s_duration	0,0028
s_country	0,0001	s_backed	0,0021	d_backers	0,0019
s_wordcount	0	s_video	0,0011	s_backed	0,0018
		s_gender	0,0008	s_video	0,0013
		s_country	0,0007	s_country	0,0009
		d_updates	0,0002	s_gender	0,0008
		s_wordcount	0,0001	s_wordcount	0,0002

Table 31: Predictors' importance computed using the function fscmrmr (from the launch to day 2)

d3		**d4**		**d5**	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,3277	d_percentpledge	0,3499	d_percentpledge	0,3438
s_category	0,0396	s_category	0,0396	s_team	0,1514
s_target	0,0383	s_target	0,0383	s_wordcount	0,0661
s_serial	0,0123	s_serial	0,013	s_category	0,0342
s_team	0,008	s_team	0,0076	s_serial	0,0235
d_updates	0,005	d_updates	0,0063	s_target	0,0148
d_comments	0,0038	d_comments	0,0037	d_comments	0,0112
s_duration	0,0024	s_duration	0,0027	d_updates	0,0089
s_backed	0,0019	d_backers	0,0024	s_backed	0,0071
d_backers	0,0016	s_backed	0,0021	s_duration	0,0069
s_video	0,0009	s_video	0,001	d_backers	0,0049
s_gender	0,0006	s_country	0,0007	s_country	0,0021
s_country	0,0005	s_gender	0,0006	s_gender	0,002
s_wordcount	0,0001	s_wordcount	0,0001	s_video	0,0019

Table 32: Predictors' importance computed using the function fscmrmr (day 3 to day 5)

d6		d7		d8	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,3114	d_percentpledge	0,3279	d_percentpledge	0,3828
s_wordcount	0,0008	s_wordcount	0,0008	s_category	0,0396
s_backed	0,0004	s_team	0,0004	s_target	0,0383
s_category	0,0001	s_category	0,0003	s_serial	0,0119
s_team	0,0001	s_serial	0,0003	s_team	0,0066
s_serial	0,0001	s_target	0,0002	d_comments	0,0043
s_country	0	d_comments	0,0002	d_updates	0,0037
s_gender	0	s_backed	0,0001	s_duration	0,0026
s_video	0	s_duration	0,0001	d_backers	0,0023
s_duration	0	d_backers	0,0001	s_backed	0,0021
s_target	0	d_updates	0,0001	s_video	0,0009
d_backers	0	s_country	0	s_gender	0,0007
d_updates	0	s_gender	0	s_country	0,0005
d_comments	0	s_video	0	s_wordcount	0,0001

Table 33: Predictors' importance computed using the function fscmrmr (day 6 to day 8)

d9		d10		d11	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,3982	d_percentpledge	0,3467	d_percentpledge	0,3719
s_category	0,0396	s_wordcount	0,0008	s_category	0,0396
s_target	0,0383	s_team	0,0004	s_target	0,0383
s_serial	0,0118	s_category	0,0003	s_serial	0,0116
s_team	0,0066	s_serial	0,0002	d_updates	0,0075
d_updates	0,0047	s_target	0,0001	s_team	0,0069
d_comments	0,0042	d_updates	0,0001	d_comments	0,0037
s_duration	0,0022	d_comments	0,0001	s_duration	0,002
s_backed	0,0015	s_country	0	s_backed	0,0015
d_backers	0,0015	s_gender	0	d_backers	0,0009
s_video	0,0008	s_video	0	s_video	0,0007
s_gender	0,0006	s_backed	0	s_gender	0,0006
s_country	0,0004	s_duration	0	s_country	0,0004
s_wordcount	0,0001	d_backers	0	s_wordcount	0,0001

Table 34: Predictors' importance computed using the function fscmrmr (day 9 to day 11)

d12		d13		d14	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,3858	d_percentpledge	0,3877	d_percentpledge	0,3815
s_category	0,0396	s_category	0,0396	s_category	0,0396
s_wordcount	0,0008	s_target	0,0383	s_target	0,0383
s_team	0,0003	s_serial	0,0118	s_serial	0,0114
s_serial	0,0002	s_team	0,0067	s_team	0,0055
s_target	0,0001	d_comments	0,0038	d_comments	0,0033
d_updates	0,0001	d_updates	0,0032	d_updates	0,0026
d_comments	0,0001	s_duration	0,0028	s_duration	0,0021
s_country	0	d_backers	0,0023	s_backed	0,0012
s_gender	0	s_backed	0,0017	s_video	0,0008
s_video	0	s_video	0,0011	d_backers	0,0008
s_backed	0	s_gender	0,0009	s_gender	0,0007
s_duration	0	s_country	0,0006	s_country	0,0003
d_backers	0	s_wordcount	0,0001	s_wordcount	0,0001

Table 35: Predictors' importance computed using the function fscmrmr (day 12 to day 14)

d15		d16		d17	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,4018	d_percentpledge	0,4345	d_percentpledge	0,4279
s_category	0,0396	s_category	0,0396	s_category	0,0396
s_target	0,0383	s_target	0,0383	s_target	0,0383
s_serial	0,0108	s_serial	0,0103	s_serial	0,0103
s_team	0,0056	s_team	0,0055	s_team	0,0055
d_updates	0,004	d_comments	0,0033	d_updates	0,003
d_comments	0,0034	d_updates	0,0032	d_comments	0,003
s_duration	0,0018	s_duration	0,0025	s_duration	0,0028
s_backed	0,0011	s_backed	0,0016	s_backed	0,0015
s_video	0,0006	s_video	0,0009	s_video	0,0009
s_gender	0,0005	s_gender	0,0007	s_gender	0,0007
d_backers	0,0005	d_backers	0,0007	d_backers	0,0006
s_country	0,0004	s_country	0,0005	s_country	0,0005
s_wordcount	0,0001	s_wordcount	0,0001	s_wordcount	0,0001

Table 36: Predictors' importance computed using the function fscmrmr (day 15 to day 17)

d18		d19		d20	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,4298	d_percentpledge	0,4279	d_percentpledge	0,4532
s_category	0,0396	s_category	0,0396	s_target	0,083
s_target	0,0383	s_target	0,0383	s_category	0,0396
s_serial	0,0102	s_serial	0,0103	s_serial	0,01
s_team	0,0055	s_team	0,0055	s_team	0,0056
d_updates	0,0038	d_updates	0,0047	d_updates	0,0054
d_comments	0,0028	d_comments	0,0029	d_comments	0,0038
s_duration	0,0017	s_duration	0,0017	s_duration	0,0018
s_backed	0,0011	s_backed	0,0011	s_backed	0,0011
s_video	0,0006	s_video	0,0006	s_video	0,0006
s_gender	0,0005	s_gender	0,0005	s_gender	0,0005
d_backers	0,0004	s_country	0,0004	d_backers	0,0005
s_country	0,0003	d_backers	0,0004	s_country	0,0002
s_wordcount	0,0001	s_wordcount	0,0001	s_wordcount	0,0001

Table 37: Predictors' importance computed using the function fscmrmr (day 18 to day 20)

d21		d22		d23	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,4817	d_percentpledge	0,4432	d_percentpledge	0,4461
s_category	0,0396	s_category	0,0396	s_category	0,0396
s_target	0,0383	s_target	0,0383	s_target	0,0383
s_serial	0,0094	s_serial	0,01	s_serial	0,0096
s_team	0,0056	s_team	0,0062	s_team	0,006
d_updates	0,0041	d_comments	0,0038	d_updates	0,0059
d_comments	0,0033	d_updates	0,0032	d_comments	0,0035
s_duration	0,0019	s_duration	0,0031	s_duration	0,0021
s_backed	0,0011	s_backed	0,0018	s_backed	0,0013
d_backers	0,0007	d_backers	0,0017	d_backers	0,001
s_gender	0,0006	s_video	0,0011	s_video	0,0007
s_video	0,0006	s_gender	0,0009	s_gender	0,0006
s_country	0,0005	s_country	0,0008	s_country	0,0004
s_wordcount	0,0001	s_wordcount	0,0001	s_wordcount	0,0001

Table 38: Predictors' importance computed using the function fscmrmr (day 21 to day 23)

d24		d25		d26	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,4461	d_percentpledge	0,4714	d_percentpledge	0,4669
s_category	0,0396	s_category	0,0396	s_category	0,0396
s_target	0,0383	s_target	0,0383	s_target	0,0383
s_serial	0,0094	s_serial	0,0095	s_serial	0,0091
s_team	0,0051	s_team	0,0056	s_team	0,0049
d_updates	0,0046	d_updates	0,0052	d_updates	0,0043
d_comments	0,0025	d_comments	0,0026	d_comments	0,0024
s_duration	0,0015	s_duration	0,0015	s_duration	0,0013
s_backed	0,001	s_backed	0,001	s_backed	0,001
s_video	0,0006	d_backers	0,0007	s_video	0,0006
d_backers	0,0006	s_video	0,0006	d_backers	0,0005
s_gender	0,0005	s_country	0,0005	s_country	0,0004
s_country	0,0004	s_gender	0,0005	s_gender	0,0004
s_wordcount	0,0001	s_wordcount	0,0001	s_wordcount	0,0001

Table 39: Predictors' importance computed using the function fscmrmr (day 24 to day 26)

d27		d28		d29	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,4714	d_percentpledge	0,4714	d_percentpledge	0,4714
s_category	0,0396	s_category	0,0396	s_category	0,0396
s_target	0,0383	s_target	0,0383	s_target	0,0383
s_serial	0,0092	s_serial	0,0093	s_serial	0,0093
d_updates	0,0058	d_updates	0,0059	s_team	0,0047
s_team	0,0053	s_team	0,0055	d_updates	0,0042
d_comments	0,0021	d_comments	0,0024	d_comments	0,0024
s_duration	0,0013	s_duration	0,0012	s_duration	0,0014
s_backed	0,0009	s_backed	0,001	s_backed	0,001
d_backers	0,0006	d_backers	0,0006	s_video	0,0005
s_video	0,0005	s_gender	0,0005	d_backers	0,0005
s_country	0,0004	s_video	0,0005	s_gender	0,0004
s_gender	0,0004	s_country	0,0003	s_country	0,0003
s_wordcount	0,0001	s_wordcount	0,0001	s_wordcount	0,0001

Table 40: Predictors' importance computed using the function fscmrmr (day 27 to day 29)

d30		**d31**	
Predictor	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,5179	d_percentpledge	0,5347
s_category	0,0396	s_category	0,0396
s_target	0,0383	s_target	0,0383
s_serial	0,0087	s_serial	0,0086
s_team	0,0045	s_team	0,0048
d_updates	0,0035	d_updates	0,0037
d_comments	0,002	d_comments	0,0021
s_duration	0,0012	s_duration	0,001
s_backed	0,0009	s_backed	0,0009
d_backers	0,0005	s_video	0,0005
s_gender	0,0004	d_backers	0,0005
s_video	0,0004	s_gender	0,0004
s_country	0,0003	s_country	0,0003
s_wordcount	0,0001	s_wordcount	0,0001

Table 41: Predictors' importance computed using the function fscmrmr (day 30 to day 31)

In figure (54), we charted the importance of the 10 static predictors considering, indeed, the static dataset (day 0), using a vertical bar chart. The graph shows, on the x-axis, the name of the 10 static predictors and on the y-axis, the predictor importance score. From the picture we can see that the most relevant static predictors are *s_serial* and *s_team,* two fundraisers-related factors and *s_target* and *s_category* which are campaign-related factors.

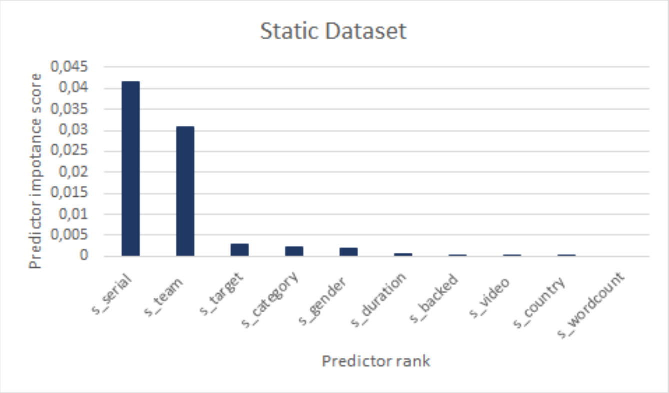


Figure 54: Relative importance of static predictors computed using the fscmrmr function

We also plotted the features' importance from the dynamic dataset at day 31 (the last considered day). Looking at the vertical bar chart, we can see that the most important predictor is a dynamic one, namely *d_percentplede,* which is the most relevant one for all the dynamic datasets considered. The most relevant static variables are *s_category, s_target, s_serial and s_team,* consistent with the results obtained in the static dataset. Hower in the dynamic one, their relative importance turns out to be lower, since there is the variable *d_percentpledge* accounting for the majority of the importance.

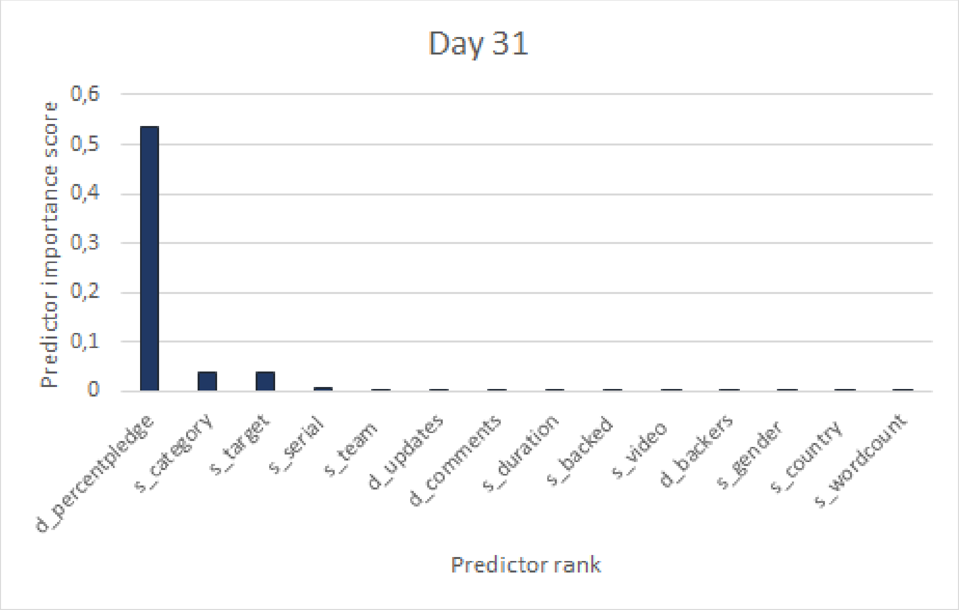


Figure 55: Relative importance of static and dynamic predictors computed using the fscmrmr function in the 31st dynamic dataset

6.3.3.2. Wrapped feature selection

In the following paragraph we will explain how we analysed the relative importance of our predictors using the *Wrapped method*. Wrapped methods implement a search in the space of the features guided by the learning algorithm. In general, the accuracy of the resulting model of the learning algorithm is calculated each time a feature is added or deleted from the set.

The debate on the use of *Wrapped* or *Filter* methods to obtain an optimal feature ranking is still alive: the method *Wrapped* generally achieves better predictive performances, especially when the optimal subset of feature is tied to a specific learning algorithm. However, the accuracy of cross-validation has a high variance on small training sets and, therefore, on such sets, W*rapped* methods tend to incur in overfitting.

So, although the *Filter* method may be more suitable for our model, given the size of our set of data, anyway, we performed this *Wrapped* analysis as a comparison of the results obtained with the *Filter* method, to be more complete in our conclusion in accessing the relative importance of the features selected. We decided to perform such method only for the two best classifiers of our model, namely for a Tree model, *Coarse Tree*, and an Ensemble algorithm, specifically *Ensemble RUSBoosted tree.*

To assess the importance of the predictors for each classification model, we used the function *predictorImportance*[29] provided by MATLAB. We have computed predictors' importance for the two classification models exported (*Coarse Tree, RUSBoosted Trees*) and for each dataset. This function, already mentioned, receives as input the classification tree previously created, and it provides as output an array with the relative score of each predictor used by the classifier (10 for the static dataset, 14 for the dynamic datasets).

The function *predictorImportance* works as follows:

imp = predictorImportance(tree)

A value equal to 0 signifies the smallest importance. This function estimates predictors' importance by *"summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes"* (MATLAB, Help Center[30]). This function receives in input the trained tree, and returns a row vector with the same number of elements as the number of predictors (columns) in the tree.

[29] Estimates of predictor importance for classification tree - MATLAB - MathWorks Italia

[30] Estimates of predictor importance for classification tree - MATLAB - MathWorks Italia

In the following subparagraphs, we show the results obtained for both the *Coarse Tree* and the *Ensemble RUSBoosted* in terms of scores associated to each predictor. The relative importance is showed in tables (from 42 to 63), where the values of the dynamic variables are highlighted in light blue.

Coarse Tree

For the *Coarse Tree*, in the following tables, we can see that consistently with previously obtained results, the most important predictors are *s_category* and *s_target* that almost for all days appear in the ranking, and *d_percentpledge,* that as the *Filter* method has shown, after day 1 seems to be the most important feature to predict overfunding. Moreover, also the other dynamic variables, such as *d_updates*, *d_backers* and *d_comments,* have a strong impact on the classification algorithm. The tables report only four predictors, because as anticipated in the explanation of *Coarse Tree* (*paragraph 6.3.2.3.*), the model allows a maximum of four splits, differently from *Medium* and *Fine Tree*. So, the 4 most important predictors in the tree construction are listed below:

s		**d1**		**d2**	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
s_category	0,0112	d_percentpledge	0,0248	d_percentpledge	0,0291
s_duration	0,0063	s_category	0,0102	s_category	0,011
s_target	0,0056	s_target	0,0057	s_target	0,0082
s_wordcount	0,0034	d_backers	0,002	s_wordcount	0,0025

Table 42: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from the launch of the campaign to day 2)

d3		d4		d5	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0568	d_percentpledge	0,0956	d_percentpledge	0,052
s_category	0,013	s_category	0,0144	s_category	0,0074
s_wordcount	0,0061	s_target	0,0091	s_target	0,0039
s_target	0,0043	s_country	0	d_backers	0,0025

Table 43: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 3 to day 5)

d6		d7		d8	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0483	d_percentpledge	0,0511	d_percentpledge	0,0602
s_category	0,0091	s_category	0,0085	s_wordcount	0,0093
s_target	0,0039	d_backers	0,0045	s_category	0,0084
d_backers	0,0027	s_target	0,0039	d_backers	0,0061

Table 44: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 6 to day 8)

d9		d10		d11	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,2909	d_percentpledge	0,0494	d_percentpledge	0,054
s_country	0	s_category	0,0042	s_category	0,0042
s_category	0	s_serial	0,0042	s_serial	0,0042
s_gender	0	s_target	0,0037	s_target	0,0037

Table 45: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 9 to day 11)

d12		d13		d14	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0489	d_percentpledge	0,0612	d_percentpledge	0,0545
s_category	0,0042	s_duration	0,0043	s_category	0,0081
s_serial	0,0042	d_updates	0,0033	s_target	0,0062
s_target	0,0037	s_category	0,0027	s_wordcount	0,0055

Table 46: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 12 to day 14)

d15		d16		d17	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0632	d_percentpledge	0,0614	d_percentpledge	0,0614
s_duration	0,0056	s_target	0,0086	s_target	0,0086
d_comments	0,0036	s_category	0,0079	s_category	0,0069
d_updates	0,0033	s_wordcount	0,0061	s_wordcount	0,0061

Table 47: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 15 to day 17)

d18		d19		d20	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0614	d_percentpledge	0,1023	d_percentpledge	0,1189
s_category	0,0069	s_category	0,0115	s_category	0,0092
s_wordcount	0,0061	s_wordcount	0,0102	s_country	0
s_target	0,006	s_country	0	s_gender	0

Table 48: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 18 to day 20)

d21		d22		d23	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,1256	d_percentpledge	0,0954	d_percentpledge	0,0936
s_category	0,0078	s_category	0,0137	s_category	0,0114
s_country	0	s_country	0	d_comments	0,0041
s_gender	0	s_gender	0	s_country	0

Table 49: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 21 to day 23)

d24		d25		d26	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,096	d_percentpledge	0,1298	d_percentpledge	0,0979
d_comments	0,0095	s_category	0,0099	d_comments	0,011
s_wordcount	0,0076	s_country	0	s_wordcount	0,0041
s_country	0	s_gender	0	s_country	0

Table 50: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 24 to day 26)

d27		d28		d29	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0997	d_percentpledge	0,1001	d_percentpledge	0,3999
d_comments	0,0092	d_comments	0,0089	s_country	0
s_wordcount	0,0042	s_wordcount	0,0041	s_category	0
s_country	0	s_country	0	s_gender	0

Table 51: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 27 to day 29)

d30		d31	
Predictor	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,138	d_percentpledge	0,4169
d_backers	0,0055	s_country	0
s_country	0	s_category	0
s_category	0	s_gender	0

Table 52: Predictors' importance computed using the function predictorImportance for the Coarse Tree (from day 30 to day 31)

Ensemble RUSBoosted Tree

Consistently with the *Coarse Tree* classifier, we listed in the tables below, only the four most important predictors, that emerged using a *Wrapped* method, for each dataset. According to the results previously obtained, the campaign-related features *s_target* and *s_category* always account for a good importance. Moreover, differently from *Coarse Tree,* the dynamic variable *d_percentpledge* confirms his domain starting from day 2 instead of day 1, but still remains the most important predictor among all the rest of the days. Nevertheless, considering this classifier, also the number of backers, *d_backers*, seems to be an important variable, ranked in the top positions for several days among the 31.

s		d1		d2	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
s_target	0,0141	s_category	0,0151	d_percentpledge	0,0188
s_category	0,0132	s_target	0,0126	s_category	0,0155
s_wordcount	0,0114	d_percentpledge	0,0104	s_wordcount	0,0135
s_duration	0,0077	d_backers	0,0093	s_target	0,0122

Table 53: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from the launch of the campaign to day 2)

d3		d4		d5	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0316	d_percentpledge	0,0349	d_percentpledge	0,0265
s_category	0,0214	s_category	0,0253	s_category	0,0225
s_duration	0,0149	s_duration	0,0242	s_target	0,0224
s_target	0,014	s_wordcount	0,0117	s_duration	0,0085

Table 54: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 3 to day 5)

d6		d7		d8	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0228	d_percentpledge	0,0264	d_percentpledge	0,0405
s_category	0,0154	s_category	0,0187	d_comments	0,0158
s_target	0,0129	s_duration	0,0117	d_backers	0,0149
d_backers	0,0115	d_backers	0,0114	s_category	0,0135

Table 55: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 6 to day 8)

d9		d10		d11	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0311	d_percentpledge	0,0278	d_percentpledge	0,0318
s_category	0,0176	s_category	0,0207	s_category	0,027
s_wordcount	0,0152	s_duration	0,013	s_target	0,0162
s_duration	0,0125	s_target	0,0117	s_wordcount	0,0084

Table 56: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 9 to day 11)

d12		d13		d14	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0242	d_percentpledge	0,0262	d_percentpledge	0,0309
s_category	0,0152	s_category	0,0155	s_category	0,0188
s_duration	0,0122	s_wordcount	0,0114	s_target	0,0127

s_target	0,0112	s_duration	0,0108	s_wordcount	0,0126

Table 57: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 12 to day 14)

d15		**d16**		**d17**	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,035	d_percentpledge	0,0502	d_percentpledge	0,0473
s_category	0,0193	s_category	0,0315	s_category	0,0234
s_target	0,0123	s_target	0,0156	s_target	0,017
d_comments	0,0102	d_comments	0,0123	s_serial	0,0125

Table 58: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 15 to day 17)

d18		**d19**		**d20**	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0473	d_percentpledge	0,0487	d_percentpledge	0,0984
s_category	0,0247	s_category	0,0227	s_category	0,0173
s_serial	0,0123	s_duration	0,0147	s_duration	0,0173
s_video	0,0101	d_backers	0,0125	d_comments	0,0129

Table 59: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 18 to day 20)

d21		**d22**		**d23**	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0605	d_percentpledge	0,1297	d_percentpledge	0,1412
s_category	0,0289	d_comments	0,0234	s_wordcount	0,0273
s_wordcount	0,024	s_wordcount	0,0198	s_duration	0,0214
d_comments	0,0175	s_category	0,0075	d_comments	0,0163

Table 60: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 21 to day 23)

d24		d25		d26	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,1568	d_percentpledge	0,0935	d_percentpledge	0,1514
d_comments	0,0156	s_category	0,017	d_comments	0,017
s_wordcount	0,012	d_backers	0,0151	s_wordcount	0,0064
s_category	0,0051	s_backed	0,0106	s_country	0

Table 61: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 24 to day 26)

d27		d28		d29	
Predictor	*Importance*	*Predictor*	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,1546	d_percentpledge	0,1551	d_percentpledge	0,093
d_comments	0,0142	d_comments	0,0137	d_backers	0,0549
s_wordcount	0,0064	s_wordcount	0,0064	s_category	0,0237
s_country	0	s_country	0	s_target	0,0226

Table 62: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 27 to day 29)

d30		d31	
Predictor	*Importance*	*Predictor*	*Importance*
d_percentpledge	0,0983	d_percentpledge	0,1583
d_backers	0,0635	d_updates	0,0264
s_target	0,0305	s_video	0,0139
s_category	0,0114	s_target	0,0109

Table 63: Predictors' importance computed using the function predictorImportance for the RUSBoosted Trees (from day 30 to day 31)

6.3.4. Impact of each static predictor on the accuracy of the model

In order to have a clearer view of the relative importance of the different static features, we decided to use a function of the *Classification Lerner app*, to assess how the accuracy of the model varied removing one predictor at a time. Indeed, on the app it is possible to remove features during the training phase, and we performed and reported this procedure for the three best classifiers (*Coarse Tree, RUSBoosted Trees and Bagged Trees*). In the tables below (64, 65, 66), we show how the accuracy of these models changes when one static feature is removed among the list of predictors used for the classification; therefore, we trained the model on 9 predictors instead of 10.

In all the tables, we highlighted in red the first three predictors that, when removed caused a significant decrease in the accuracy, meaning that they are relevant for the prediction.

Coarse Tree

When using all the 10 static predictors, the accuracy for the *Coarse Tree* is 70,90%. Removing the static features *s_category, s_target* and *s_gender* the accuracy of the model decreases, consistent with the results obtained with previous methods.

Excluded Predictors	Accuracy
s_category	64,50%
s_target	64,50%
s_gender	67,30%
s_wordcount	67,50%
s_serial	69,10%
s_team	69,10%
s_backed	70,00%
s_duration	70,00%
s_country	70,90%
s_video	70,90%

Table 64: How the accuracy changes when one static predictor is not considered for training the Coarse Tree

Ensemble Bagged Tree

For the *Ensemble Bagged Tree*, the accuracy obtained considering all the static variables is equal to 67,30%. Removing *s_team, s_target* and *s_country* the accuracy drops, and again this result confirms the findings previously obtained.

Excluded Predictors	Accuracy
s_team	63,60%
s_target	65,50%
s_country	66,40%
s_backed	66,40%
s_gender	67,30%
s_duration	68,20%
s_video	70,00%
s_wordcount	70,00%
s_category	70,90%
s_serial	70,90%

Table 65: How the accuracy changes when one static predictor is not considered for training the Bagged tree

Ensemble RUSBoosted Tree

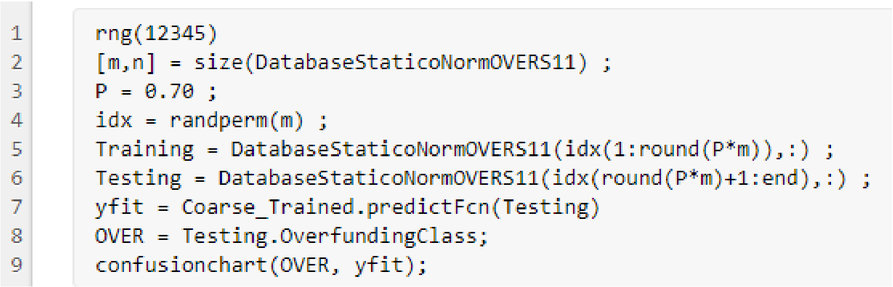
Also, in this case, consistent with the other two trees, we find that when the predictors *s_category*, *s_target and s_team* are removed, the accuracy drastically drops with respect to the accuracy when all the static predictors are included (73,60%).

Excluded Predictors	**Accuracy**
s_category	60,00%
s_target	65,50%
s_team	66,40%
s_wordcount	67,30%
s_backed	69,10%
s_country	70,90%
s_video	70,90%
s_serial	70,90%
s_gender	72,70%
s_duration	74,50%

Table 66: How the accuracy changes when one static predictor is not considered for training the RUSBoosted tree

6.3.5. Testing

In the end, the testing phase has been realised. Firstly, all the trained model of the three best classifiers were exported after the training phase. Then they have been tested with a new set of instances, the 30% of the total campaigns (47 projects), randomly chosen by the function previously implemented in MATLAB *(paragraph 6.3.2.2.)*. In particular the following lines of code were implemented:



```
rng(12345)
[m,n] = size(DatabaseStaticoNormOVERS11) ;
P = 0.70 ;
idx = randperm(m) ;
Training = DatabaseStaticoNormOVERS11(idx(1:round(P*m)),:) ;
Testing = DatabaseStaticoNormOVERS11(idx(round(P*m)+1:end),:) ;
yfit = Coarse_Trained.predictFcn(Testing)
OVER = Testing.OverfundingClass;
confusionchart(OVER, yfit);
```

The example, in the figure above, regards the trained *Coarse tree* in the static dataset. Precisely, after having exported the *Coarse tree* from the *Classification Learner app*, we obtained a "*Corse_trained*" model structure to be used to make predictions using new data. In order to adopt the exported model with a new set of data, we employed the form *"yfit"* [31]which provides as output an array containing a class prediction for each data point.

Then, we extracted from the Testing dataset the "*OverfundingClass*" column, and we created a confusion chart plotting both the "*OverfundingClass*" and the *yfit* variable, in order to understand whether our model was able to predict the true class of the instances in the testing phase.

Below an example of the confusion matrix obtained:

[31] https://www.mathworks.com/help/stats/compactregressiontree.predict.html

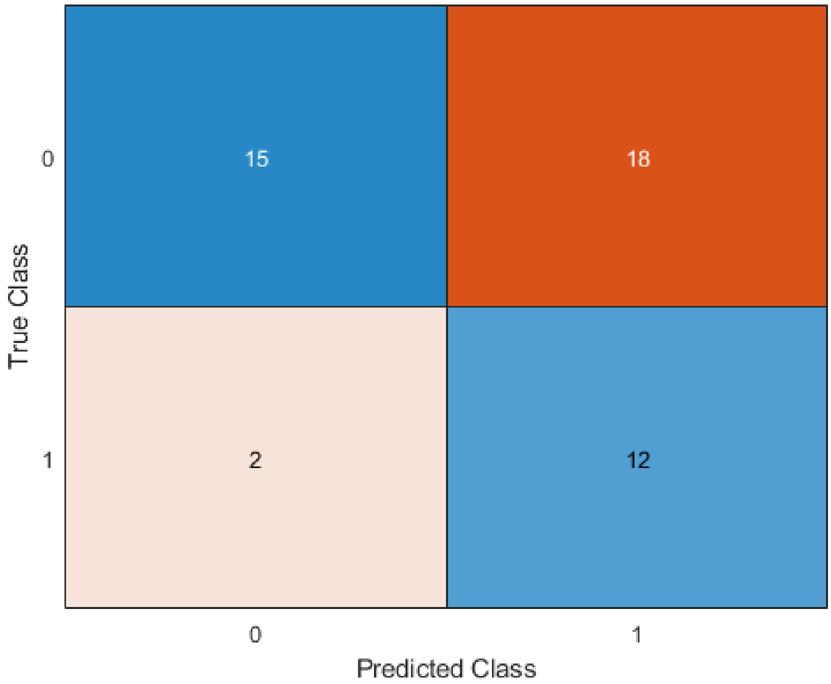


Figure 56: Confusion matrix obtained among OverfundingClass and Yfit

At this phase of the analysis, four indicators have been calculated for each dataset (1 static and 31 dynamics): *accuracy, precision, recall, F1 score.* In the following tables all the values associated to these parameters are reported, with a visual representation of the trend of the accuracy, precision, recall from the day when the campaign begins, till day 31st.

Coarse Tree

Regarding the first classifier taken into account, namely *Coarse Tree,* it reaches very high performances in all the four parameters, specifically achieving a number of *"False Negative"* equal to 0 (recall up to 100%) from day 20th to day 31st.

Dataset	***Accuracy***	***Recall***	***Precision***	***F1 Score***
s	57,45%	85,71%	40,00%	54,55%
d1	76,60%	71,43%	58,82%	64,52%
d2	89,36%	92,86%	76,47%	83,87%
d3	87,23%	92,86%	72,22%	81,25%
d4	80,85%	92,86%	61,90%	74,29%
d5	87,23%	85,71%	75,00%	80,00%
d6	82,98%	92,86%	65,00%	76,47%
d7	80,85%	85,71%	63,16%	72,73%
d8	93,62%	85,71%	92,31%	88,89%
d9	87,23%	92,86%	72,22%	81,25%
d10	82,98%	92,86%	65,00%	76,47%
d11	82,98%	92,86%	65,00%	76,47%
d12	82,98%	92,86%	65,00%	76,47%
d13	91,49%	85,71%	85,71%	85,71%
d14	95,74%	92,86%	92,86%	92,86%
d15	93,62%	92,86%	86,67%	89,66%
d16	93,62%	85,71%	92,31%	88,89%
d17	93,62%	85,71%	92,31%	88,89%
d18	93,62%	85,71%	92,31%	88,89%
d19	93,62%	85,71%	92,31%	88,89%
d20	82,98%	100,00%	63,64%	77,78%
d21	82,98%	100,00%	63,64%	77,78%

d22	82,98%	100,00%	63,64%	77,78%
d23	87,23%	100,00%	70,00%	82,35%
d24	87,23%	100,00%	70,00%	82,35%
d25	85,11%	100,00%	66,67%	80,00%
d26	93,62%	100,00%	82,35%	90,32%
d27	95,74%	100,00%	87,50%	93,33%
d28	95,74%	100,00%	87,50%	93,33%
d29	95,74%	100,00%	87,50%	93,33%
d30	95,74%	100,00%	87,50%	93,33%
d31	97,87%	100,00%	93,33%	96,55%

Table 67: Result of the tesing phase of the Coarse tree

The trend of the accuracy along the duration of the campaigns tends to become more linear by day 26. Overall, the curve is fairly smooth (Figure 57).

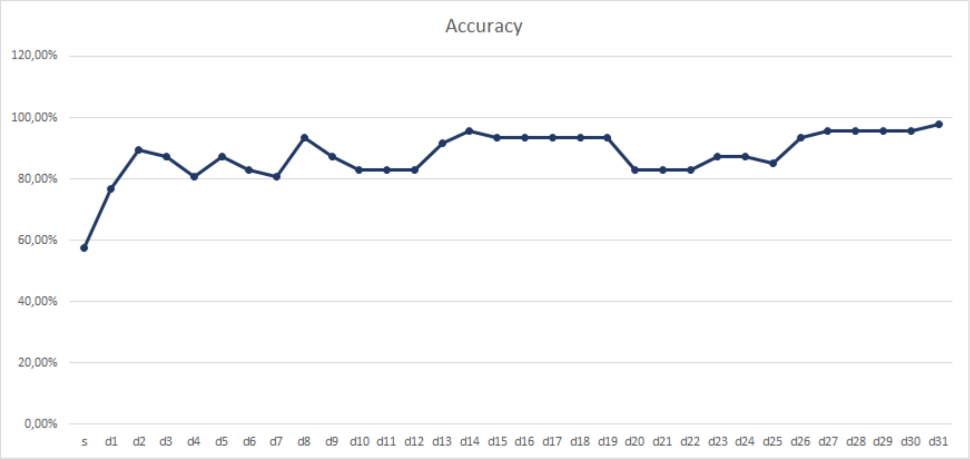


Figure 57: Coarse Tree accuracy from the launch of the campaign to day 31

As previously anticipated, the values related to the recall (Figure 58) reach the 100% at day 20th and remains linear till the end of the campaign. The model is 100% sensitive, meaning that it correctly identifies each class every time that occurs.

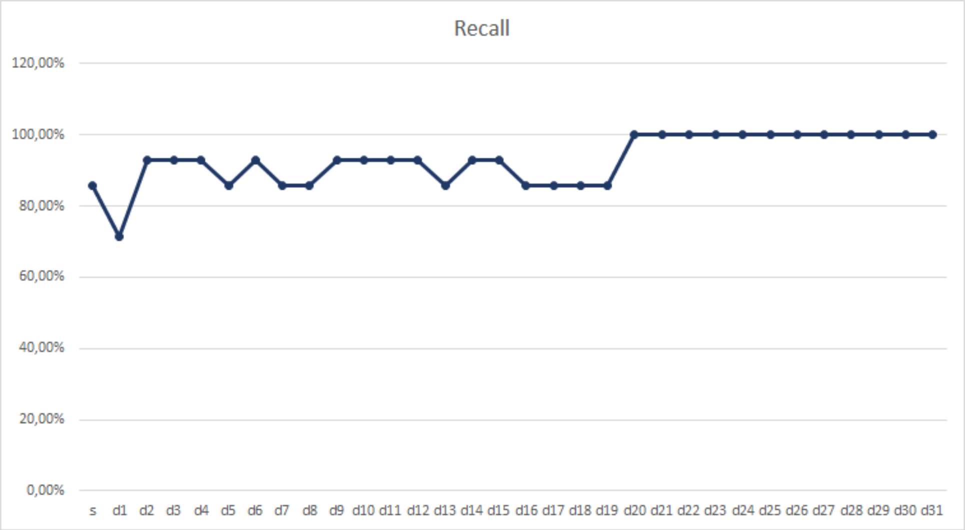


Figure 58: Coarse tree Recall from the launch of the campaign to day 31

Fortunately, the classification model is not only extremely sensitive, but also reaches very high level of precision: 93,33 % at the end of day 31st (Figure 59), therefore every time the model foresees the event, it misses very rarely.

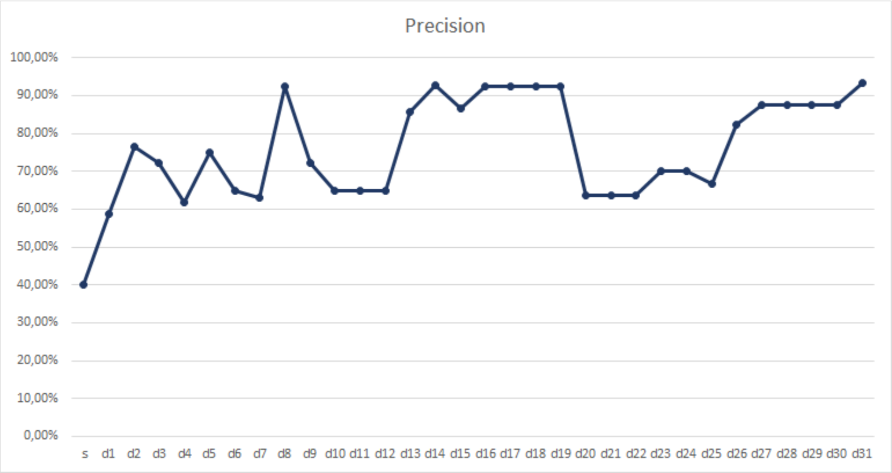


Figure 59: Coarse tree Precision from the launch of the campaign to day 31

Ensemble Bagged Tree

Consistently with the results highlighted for the Coarse tree, also the Ensemble Bagged tree achieves very good performances in all the indicators considered, reaching a recall value equal to 100% starting from day 21st. Compared to the other two trees under investigation the Ensemble Bagged Tree is the one managing to reach very high values in all the 4 parameters in the static dataset too.

Dataset	*Accuracy*	*Recall*	*Precision*	*F1 Score*
s	74,47%	50,00%	58,33%	53,85%
d1	80,85%	71,43%	66,67%	68,97%
d2	74,47%	50,00%	58,33%	53,85%
d3	87,23%	92,86%	72,22%	81,25%
d4	89,36%	92,86%	76,47%	83,87%
d5	91,49%	92,86%	81,25%	86,67%
d6	89,36%	92,86%	76,47%	83,87%
d7	89,36%	92,86%	76,47%	83,87%
d8	89,36%	92,86%	76,47%	83,87%
d9	91,49%	92,86%	81,25%	86,67%
d10	91,49%	92,86%	81,25%	86,67%
d11	91,49%	92,86%	81,25%	86,67%
d12	91,49%	92,86%	81,25%	86,67%
d13	97,87%	92,86%	100,00%	96,30%
d14	95,74%	92,86%	92,86%	92,86%
d15	97,87%	100,00%	93,33%	96,55%
d16	85,11%	85,71%	70,59%	77,42%

d17	91,49%	85,71%	85,71%	85,71%
d18	87,23%	85,71%	75,00%	80,00%
d19	91,49%	85,71%	85,71%	85,71%
d20	91,49%	92,86%	81,25%	86,67%
d21	93,62%	100,00%	82,35%	90,32%
d22	89,36%	100,00%	73,68%	84,85%
d23	89,36%	100,00%	73,68%	84,85%
d24	87,23%	100,00%	70,00%	82,35%
d25	91,49%	100,00%	77,78%	87,50%
d26	95,74%	100,00%	87,50%	93,33%
d27	95,74%	100,00%	87,50%	93,33%
d28	95,74%	100,00%	87,50%	93,33%
d29	93,62%	100,00%	82,35%	90,32%
d30	97,87%	100,00%	93,33%	96,55%
d31	97,87%	100,00%	93,33%	96,55%

Table 68: Result of the tesing phase of the Bagged tree

The trend of the accuracy tends to vary during the days, and differently from the Coarse tree, it does not reach stabilized values from one particular day on. Only days 30 and 31 achieve linear values (Figure 60).

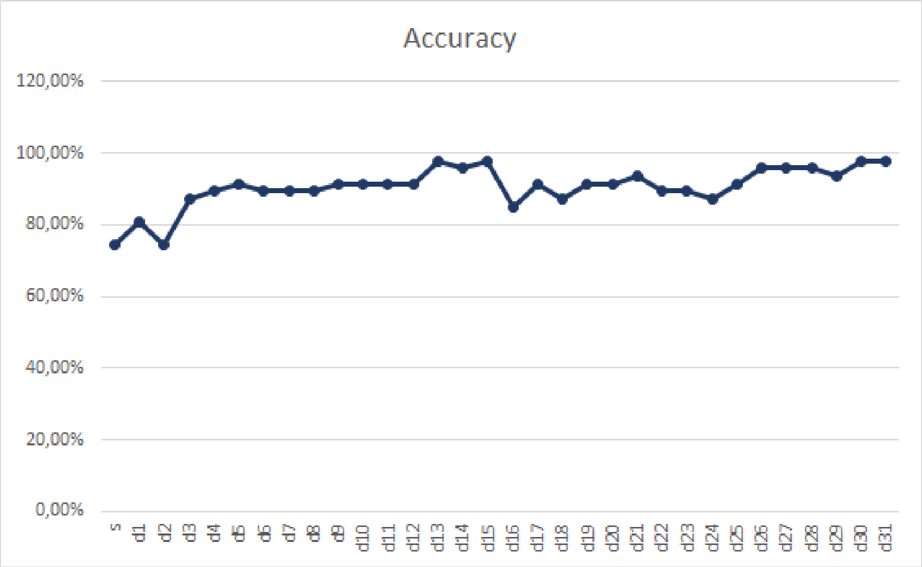


Figure 60: Bagged Tree accuracy from the launch of the campaign to day 31

As mentioned at the beginning, the recall values achieve a percentage equal to 100% starting from day 21st and remain such until the end of the campaign. With the exception of the static dataset and the first two dynamic ones, the recall indicator is always quite stable around the percentages of 80% to 100%

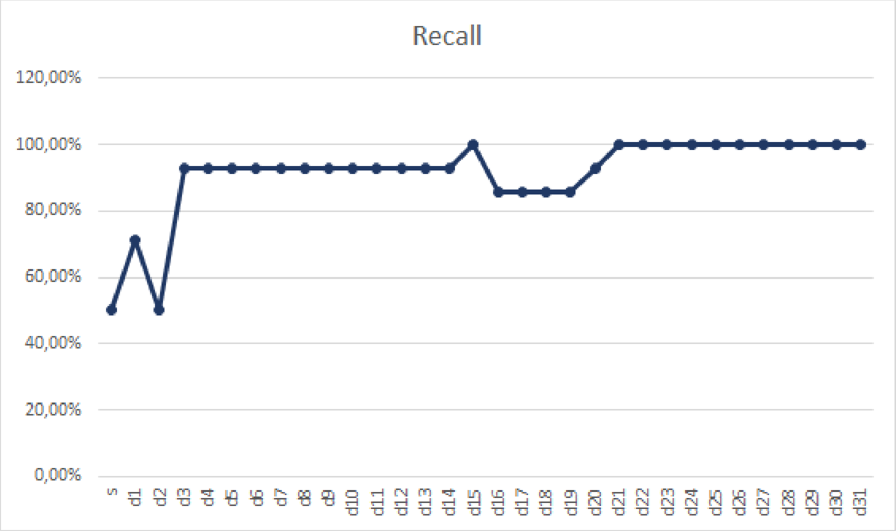


Figure 61: Bagged Tree Recall from the launch of the campaign to day 31

Also, the precision index reaches very good results, achieving a percentage of 93,33 % at the end of day 31st meaning, as in the case of the Coarse tree, that our model is highly precise in foreseeing the event under study.

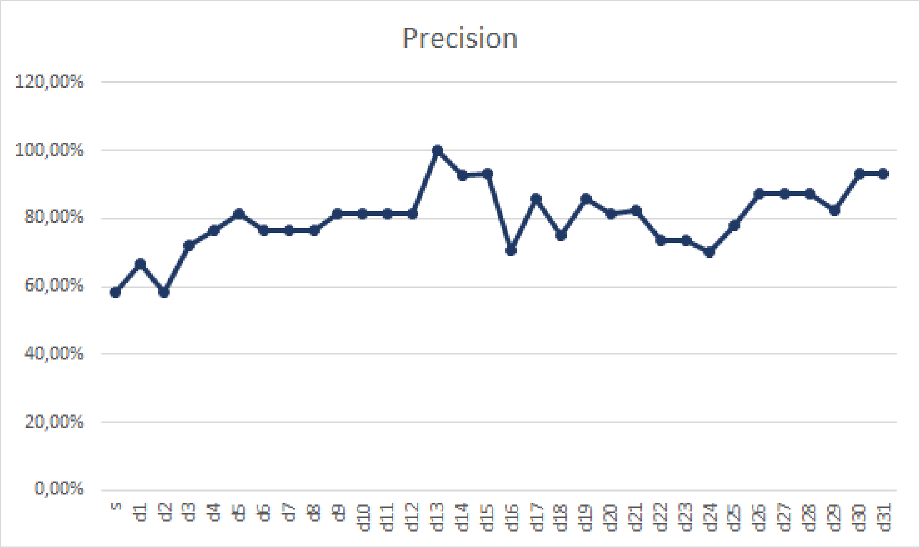


Figure 62: Bagged Tree precision from the launch of the campaign to day 31

Ensemble RUSBoosted Trees

Again, as for the *Coarse Tree* and *Ensemble Bagged*, also for this last classifier considered, the values of accuracy, recall, precision and F1 scores well performed. Precisely, the accuracy related to the static dataset, is very low compared to the one obtained at day 1 (+19,15%), similar to the results obtained for the *Coarse Tree*, confirming the importance of dynamic variables in predicting the outcomes.

Dataset	***Accuracy***	***Recall***	***Precision***	***F1 Score***
s	57,45%	71,43%	38,46%	50,00%
d1	76,60%	71,43%	58,82%	64,52%
d2	80,85%	85,71%	63,16%	72,73%
d3	78,72%	85,71%	60,00%	70,59%
d4	80,85%	92,86%	61,90%	74,29%
d5	87,23%	85,71%	75,00%	80,00%
d6	89,36%	85,71%	80,00%	82,76%
d7	89,36%	85,71%	80,00%	82,76%
d8	85,11%	92,86%	68,42%	78,79%
d9	87,23%	92,86%	72,22%	81,25%
d10	87,23%	92,86%	72,22%	81,25%
d11	87,23%	92,86%	72,22%	81,25%
d12	80,85%	92,86%	61,90%	74,29%
d13	85,11%	92,86%	68,42%	78,79%
d14	82,98%	92,86%	65,00%	76,47%
d15	85,11%	92,86%	68,42%	78,79%
d16	82,98%	85,71%	66,67%	75,00%
d17	78,72%	85,71%	60,00%	70,59%
d18	80,85%	85,71%	63,16%	72,73%
d19	82,98%	85,71%	66,67%	75,00%
d20	85,11%	100,00%	66,67%	80,00%
d21	85,11%	100,00%	66,67%	80,00%

d22	85,11%	100,00%	66,67%	80,00%
d23	93,62%	100,00%	82,35%	90,32%
d24	91,49%	100,00%	77,78%	87,50%
d25	93,62%	100,00%	82,35%	90,32%
d26	91,49%	100,00%	77,78%	87,50%
d27	93,62%	100,00%	82,35%	90,32%
d28	93,62%	100,00%	82,35%	90,32%
d29	93,62%	100,00%	82,35%	90,32%
d30	97,87%	100,00%	93,33%	96,55%
d31	97,87%	100,00%	93,33%	96,55%

Table 69: Result of the tesing phase of the RUSBoosted tree

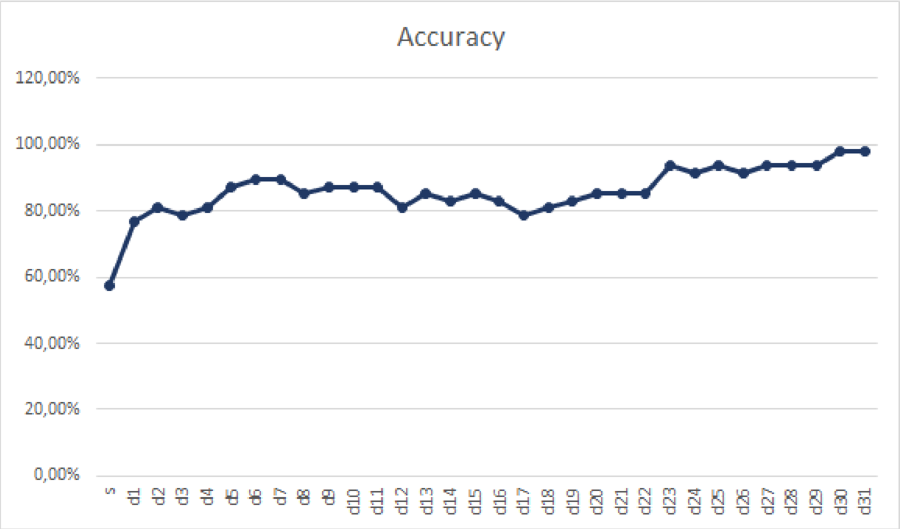


Figure 63: RUSBoosted Tree precision from the launch of the campaign to day 31

The recall of the classification model is similar to the previous Trees, settling its value at 100% from day 20th, so identifying the class every time that occurs.

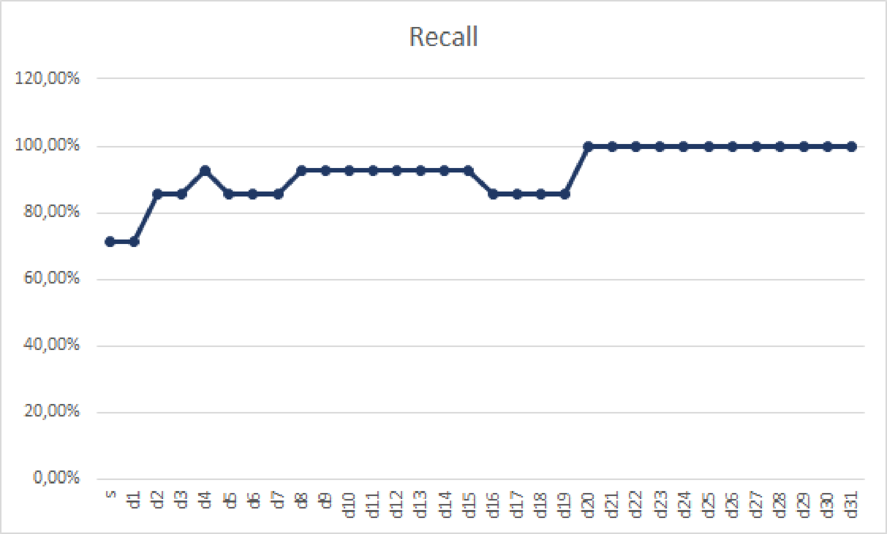


Figure 64: RUSBoosted Tree recall from the launch of the campaign to day 31

Also, the trend of the precision in quite similar to the previous shown, reaching a value of 93,33% at the end of the days.

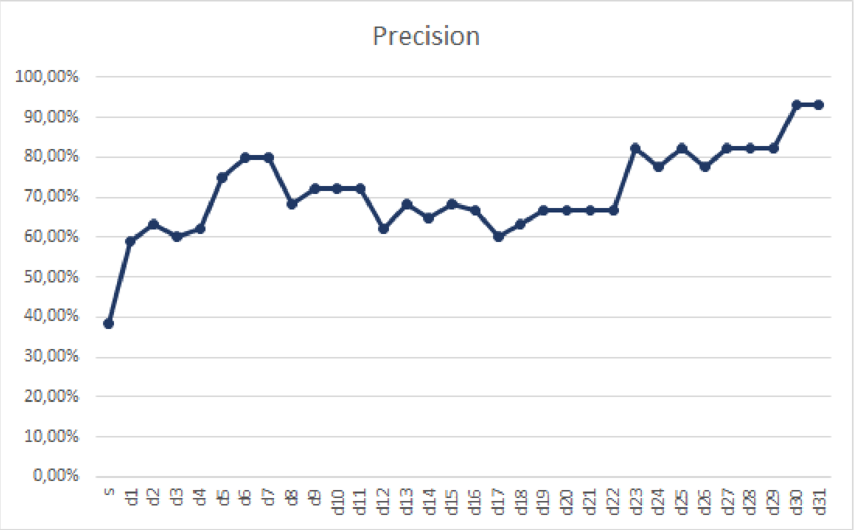


Figure 65: RUSBoosted Tree precision from the launch of the campaign to day 31

Chapter 7 – Conclusion and results

In this Chapter, we first explore deeply the results obtained in the section *“Methodology and Empirical findings”*, addressing properly the research objectives developed in the third chapter (*Chapter 3 - Research Objectives*). Secondly, we explain the limitations and further development for future researches that could be undertaken by scholars. Then, we highlight the implications of our Dissertation from both a theoretical and practical perspective. Finally, we point out the final conclusions of our research.

7.1. Results related to the first objective of the Thesis (01)

As highlighted in the *Chapter 3 "Research Objectives"*, literature has never proposed a clear definition and an operationalization able to discriminate between success and overfunding. Therefore, the first aim of this thesis was to develop an empirical classification model able, with a certain degree of confidence level, to predict the outcome of a campaign, recognising between those of success and those of *“over-success”*.

As developed in the *Methodology*, we fixed an overfunding percentage threshold, calculated as the ratio between the pledge and the target amount, among 120% and 175%, trying to understand which of these percentages allowed our classification tool to identify the phenomenon of overfunding with the highest accuracy possible. We deeply evaluated the best accuracy obtained in the Training phase varying the overfunding threshold. Below, we mapped (Table 70) as the days progress, the threshold with the highest accuracy for the different 32 datasets.

Threshold	***Frequency***
120%	0
125%	1
130%	0
135%	0
140%	0
145%	0
150%	6
155%	0
160%	0
165%	1
170%	2
175%	22

Table 70: number of times each overfunding threshold has been the best in terms of accuracy

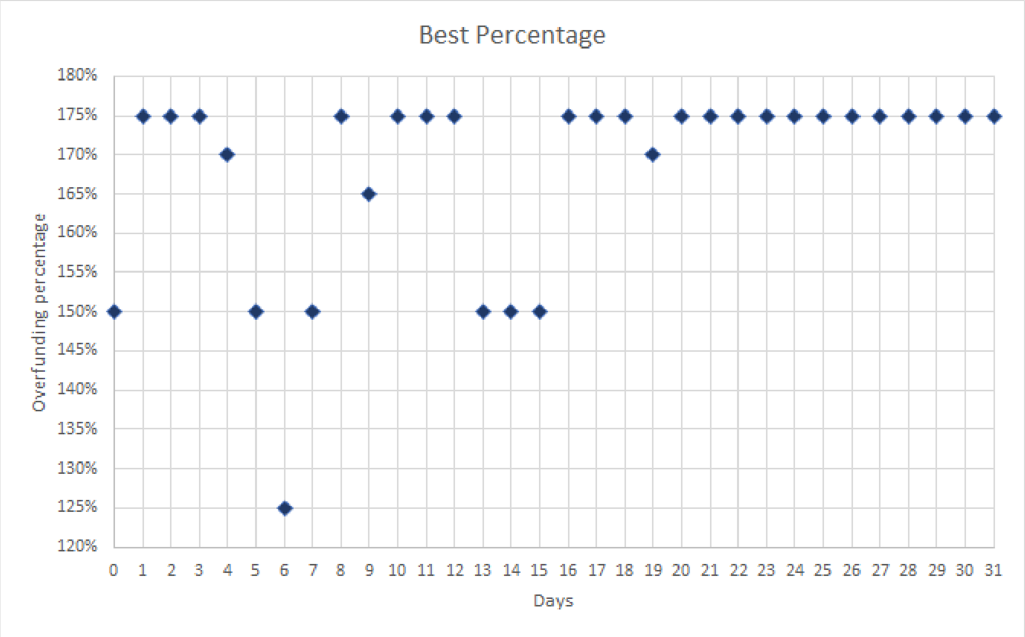


Figure 66: graphical mapping of the best overfunding percentage in terms of accuracy

Contrary to what we expected, the model works very well when the overfunding threshold is elevated. Indeed, with low percentages such as 130% and 135%, the dataset had more balanced classes with an almost equal number of successful and over-successful campaigns, so we initially believed that the classification model would have performed better using this discrimination-level. Instead, although with 175% the two discriminated classes are less quantitatively equal (but still below a 30-70 difference), the classifier more easily recognises overfunding.

Indeed, comparing the daily accuracy of the first 2 weeks, reported in a table for each of the 3 best Classifiers (67, 68, 69), those obtained with 175% perform highly better than those derived by choosing as overfunding threshold of 130%. This confirms what previously stated, namely that, although with 130% the outcomes *"success"* and *"over-success"* classes were balanced equally (48% success, 52% overfunding), the model trains better by fixing overfunding at high percentages with more unbalanced classes (67% success, 33% overfunding).

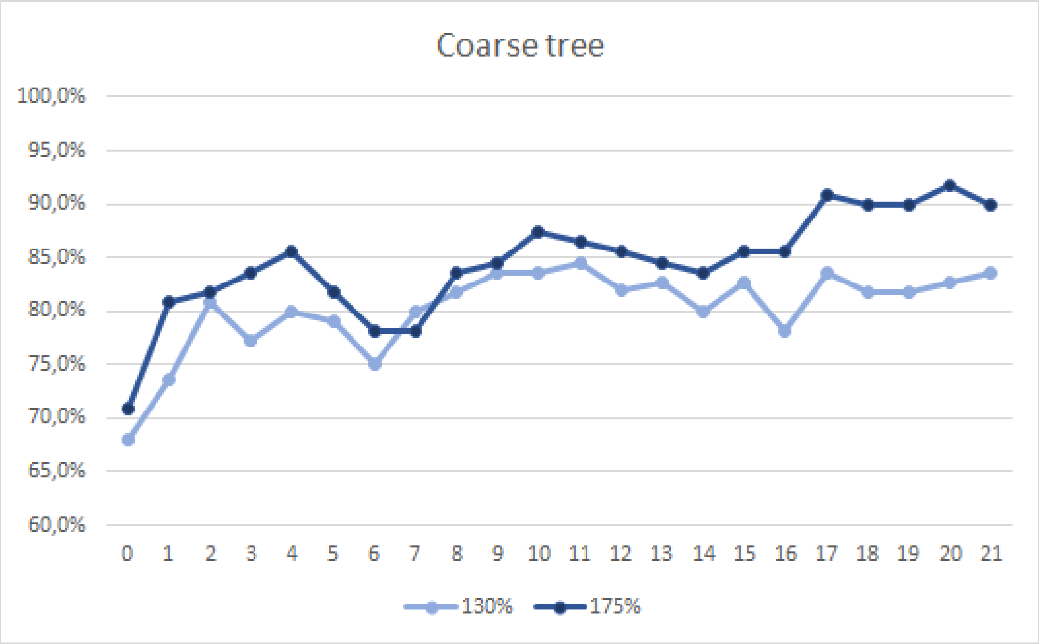


Figure 67: comparison of the accuracy reached for the Coarse tree using 130% and 175%

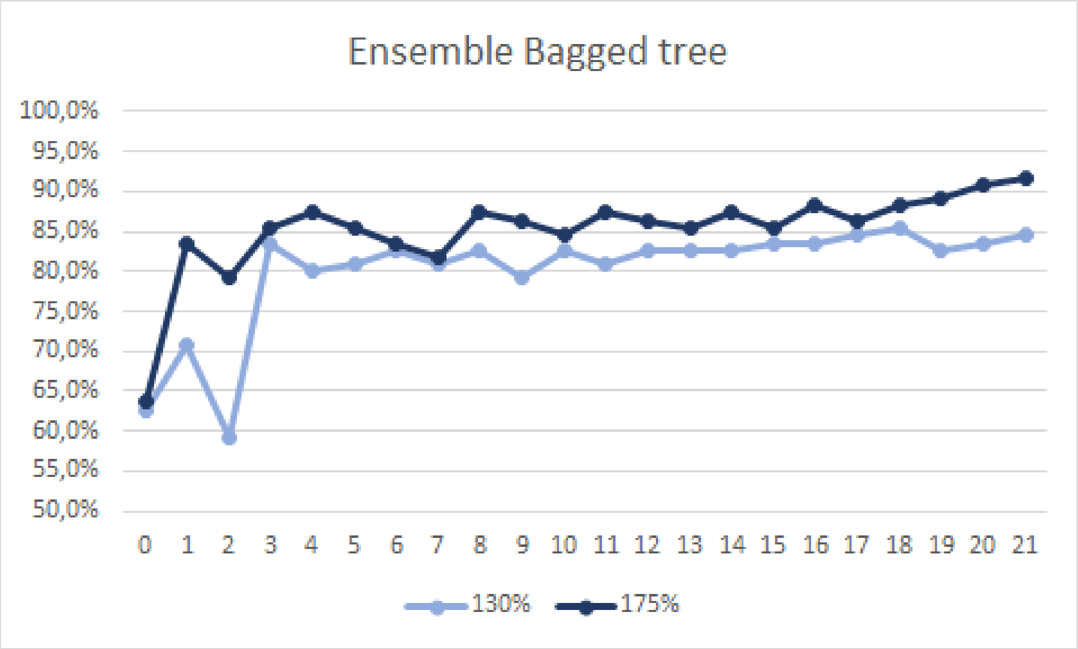


Figure 68: comparison of the accuracy reached for the Bagged tree using 130% and 175%

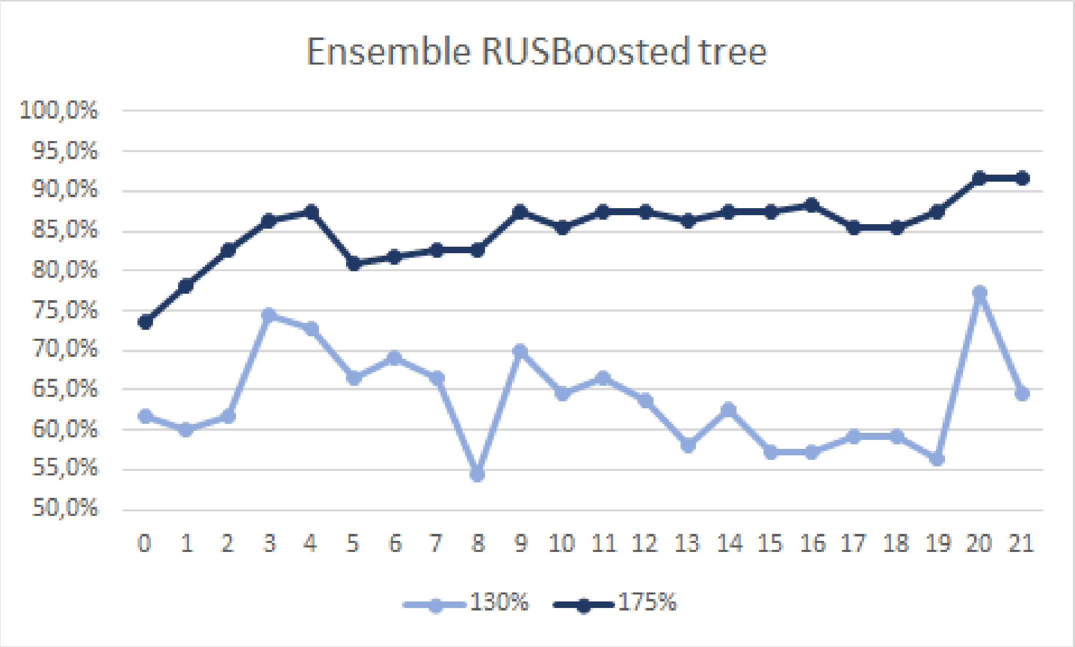


Figure 69: comparison of the accuracy reached for the RUSBoosted tree using 130% and 175%

Thanks to our analysis we were able to make a step forward with respect to extant literature, where previous scholars had left very blurred the boundaries between the two phenomena. Indeed, we demonstrated, thanks to our Classification model and the use of machine learning techniques, that success and overfunding are two clearly distinguished conditions. Specifically, our empirical model provides a possible operationalisation that could be used to define and differentiate the two phenomena, which directly emerge from the confidence level through which the classifier is able to discriminate among success and overfunding:

$$\textit{Overfunding:} \qquad \frac{Pledge}{Target} \geq 175\%$$

$$\textit{Success:} \qquad 100\% \leq \frac{Pledge}{Target} < 175\%$$

Finally, we conclude this section embracing the definition provided by Mollick (2014): *"Overfunding is especially used when a project's funding is considerably higher than its funding goal",* also managing to concretely define the general concept of *"considerably higher"* into mathematical terms.

Moreover, with our research we detach ourselves from the definition given by all Frydrych et al., 2014; Ma X. et al., 2018; Li Y. et al., 2020; Cordova et al., 2015, who state that *"a project is called overfunded the moment its funding exceeds the goal"*, since we cannot define overfunding as a simple overcoming of the target (even of a few dollars), but we can identify it when the pledge exceeds by high percentages the target (for our dataset, indeed, 175%), otherwise the campaign is simply successful but not overfunded.

7.2. Results related to the second objective of the Thesis (02)

In this paragraph we will explore the findings coming from the *Features' ranking paragraph* (*6.3.3.*), to achieve the second Objective of the thesis, namely asses which are the most relevant static and dynamic variables in predicting overfunding. According to the analysis that we followed and our willingness to highlight the impact of both static and dynamic predictors, we started reasoning about some conclusions for the ones already available at the offset of the campaign, namely the statics. As emerged by the *Filter, Wrapped* method and by assessing the accuracy of the model excluding one predictor at a time, the most relevant static variables to predict overfunding are: *s_category,* namely the typology of the project, *s_target* the funding goal set at the beginning of the campaign, s_serial, a variable representing whether the fundraiser is serial (has already launched campaigns before) or not, and lastly *s_team,* namely whether the campaign is launched by a single fundraiser or a team.

In the following table (71), we provide an overview of the comparison with the most important static variables to predict success, which derived from the primary binary analysis conducted by Annunziata and Aversa (2020), with the aim to discriminate between successful and unsuccessful campaigns.

Success		***Overfunding***	
Category	*Predictor*	*Category*	*Predictor*
Campaign - related factors	s_picture	**Campaign - related factors**	s_target
	s_target		s_category
	s_backed	**Fundraisers - related factors**	s_serial
	s_category		s_team

Table 71: Comparison of the most important static factors between success and overfunding

As highlighted in the table above, the main static factors to predict the successful outcome of a campaign are *s_picture, s_target, s_backed* and *s_category,* all campaign-related factors, meaning that for success these kinds of predictors have a stronger weight when performing a campaign's analysis. Different is what concerns the overfunding prediction, since the most relevant factors belong to two different categories. Precisely, *s_category* and *s_target* are two campaigns-related factors, while *s_serial* and *s_team* are fundraisers-related. This implies that, for the overfunding, *"human"* factors, specifically some characteristic regarding the project creators, start to weight more in the ability to predict the phenomenon. The result related to *s_serial* is consistent with what is stated in literature, in particular by Koch and colleagues 2018, that sustain that there is a strong correlation between the fundraiser being serial and the overfunding probability.

Moreover, as mentioned, the second part of our *Research Objective (O2)* proposes to deepening the importance of the dynamic predictors in the model. First of all, it is interesting to point out how the accuracy that the classification tool reaches drastically increases as the days of the campaign progress (Figure 70), reaching a level of 98,20% in day 31st. This means that the dynamic variables have a great positive impact on the capability of the model to predict the overfunding of a campaign.

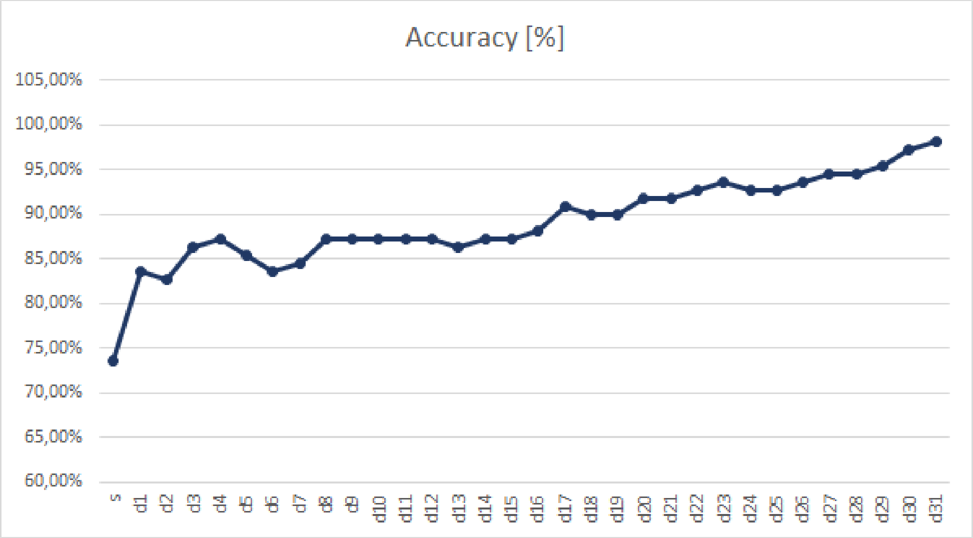


Figure 70: accuracy trend visualized over the 31 dynamic datasets

For all the days following the launch of the campaign, the importance of the static variables to predict overfunding is still elevate, as emerged by the *Wrapped* and *Filter* methodology, but their influence is much lower than the one of the dynamic variable *d_percentpledge*, representing the sum of the pledged money received by the campaign up to day *t* over the target goal (*s_target*). The importance of this factor is consistent with the overfunding definition we provided in the previous paragraph, since it is strongly related to the ability of the campaign to reach a ratio between pledge and funding target higher that 175%. For this reason, a measure that accounts for this progress is useful for the classifiers to predict campaigns' overfunding.

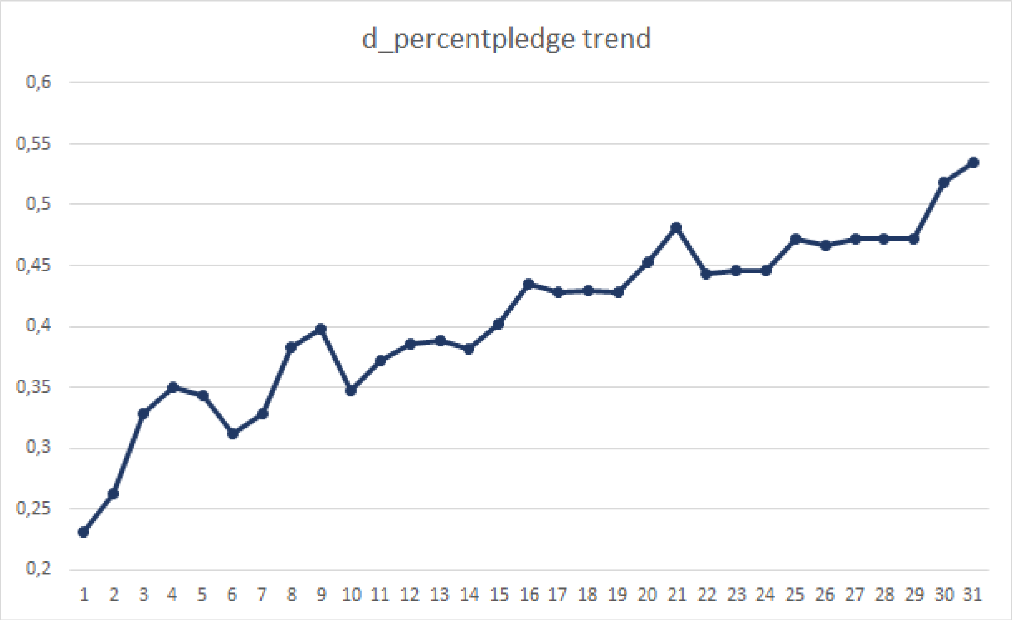


Figure 71: trend of the relative importance of the variable d_percentpledge visualized over the 31 dynamic datasets

Moreover, as it can be easily visualized in the picture above (71), the value of the variable *d_percentpledge* tends to increase during the campaign duration, becoming strongly relevant in the last days. This result is consistent with many literature findings, in particular those stating that backers prefer to invest in already pledged campaigns, since in this way the risk and uncertainty associated with the investment is reduced (Theerthaana and Manzoor, 2019). Indeed, as the days pass, the amount pledged to the campaign increases, and this will attract new funders. This is the reason why this variable is so important in predicting overfunding and also why its relative importance increases during the days.

7.3. Results related to the third objective of the Thesis (03)

As reported in *Chapter 3,* the literature on overfunding has not yet focused on understanding whether, as in the case of success, it is sufficient to consider only the first week as the most critical period, or it is needed to take into account the entire lifespan of the campaign to have a complete overview of the phenomenon. Therefore, to find out a proper insight about the impact of *"time"* in overfunding, we developed the empirical model considering the entire life-span of the majority of our campaigns in the datasets (31 days). After the analysis carried out, we could finally state that, in order to have a complete spectrum of overfunding, it is necessary to look at the entire duration of the campaign, in our case 31 days. In fact, on the 31st day, all the campaigns of the sample reach overfunding and our model achieves very high values of accuracy in discriminating between success and overfunding.

However, it is worthy to underline, that for our sample of data, more than 50% of the campaigns, that at the end of the entire duration should have reached overfunding, has already achieved it by the seventh day. This result is completely in line with the findings about success, meaning that both in the cases the first week may be considered the most insightful to obtain a successful and/or o*ver-successful* outcome. In the figure below (72), we show the percentage of campaigns achieving overfunding (in relation to the total number that had to reach overfunding at the end of the period) during the first 4 campaign weeks.

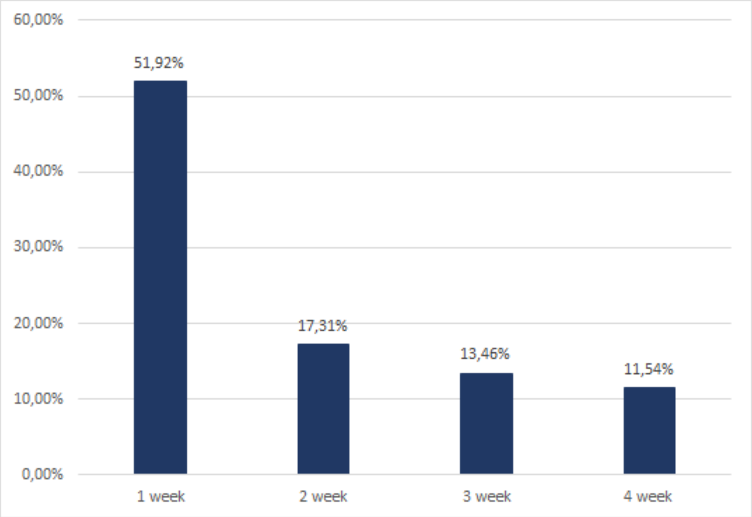


Figure 72: Percentage of campaigns reaching overfunding in week 1, 2, 3, 4

Moreover, from the following figure (73), it is easily highlighted that at the end of the 4th week of the life-cycle, the classification model predicts overfunding with an accuracy of almost 99%. This confirms that the best outcome possible in terms of the accuracy of the prediction is achievable with a broad spectrum of 31 days. However, it is possible to predict the overfunding phenomenon with a more than acceptable level of accuracy at the end of the first seven days.

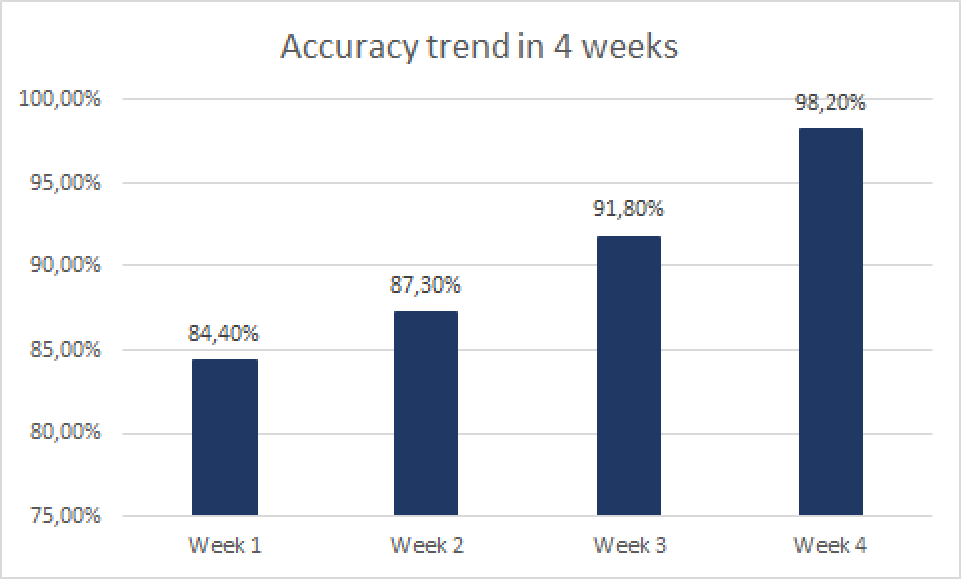


Figure 73: trend of the percentage of the accuracy of the classification model during the 4 campaigns' week

All these reasonings led us to conclude that, even if the best outcome possible implementing a classification tool is achievable considering the entire life-cycle of the campaign, in case a comparison between success and overfunding has to be done, practitioners could focus on the first seven days of the week. Indeed, as previously demonstrated by Annunziata and Aversa (2020), an empirical model predicting success reaches at the end of the 7th day an accuracy of 90,7%, instead the classification tool implemented to discriminate between success and overfunding ends up with an accuracy of 84,40% at the end of the first week.

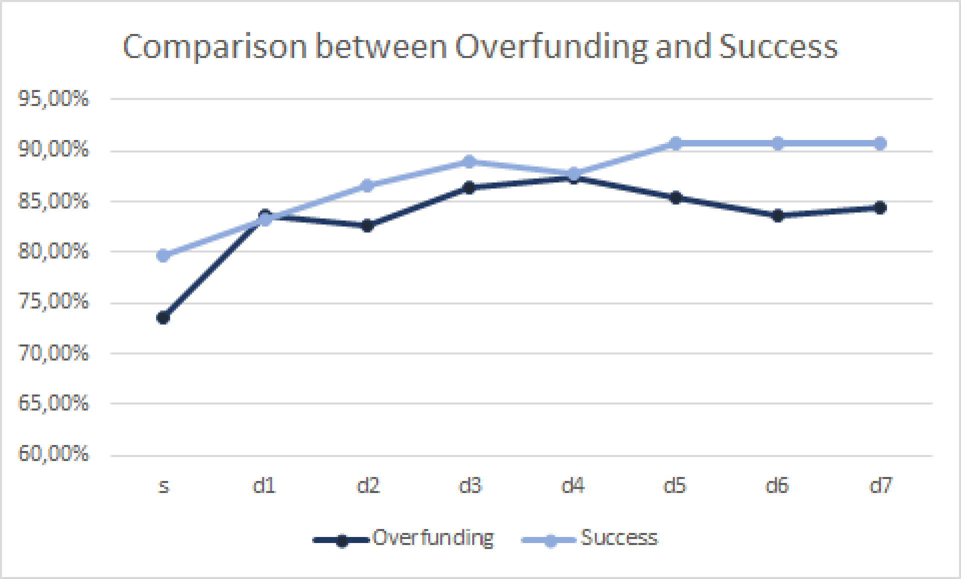


Figure 74: comparison between success and overfunding of the accuracy reached during the first 7 campaign days

7.4. Limitations and further research

The methodology carried out to achieve the research objectives presents several limitations, which leave ample room for future researches on overfunding, a very recent theme that is starting to raise scholars' attention among the still unexplored areas of crowdfunding.

As pointed out while discussing the "*Methodology and Empirical Findings" paragraph*, two distinctive binary analysis have been conducted (resuming the work of Annunziata and Aversa, 2020) to build a model able to discriminate first between success and unsuccess and then, at a later stage, between successful and "*over-successful"* outcomes. In a preliminary phase of our research, we tried to model our analysis as a multi-class problem. The unique model built should have returned 3 different outputs: 0 associated to negative outcomes of the campaign (if in their life-cycle, they have not been able to reach the target amount), 1 to successful campaign, 2 when the project might have been considered *"overfunded"*. With this procedure it would have been possible to use the entire database consisting of 352 campaigns. However, with this process, the classes imported into the *Classification Learner app* were not balanced, since 195 campaigns were unsuccessful, and the remaining 157 divided between success and overfunding. Indeed, our sample is not evenly distributed between successful and not successful campaigns. The presence of a higher number of unsuccessful campaigns, compared to the successful ones led to some biases in the results: when we trained the algorithm, the results in term of accuracy were very low (below 60%). Due to this outcome, we decided to divide the analysis into two parts, and to keep only the 157 campaigns to discriminate between success and overfunding. Therefore, future analysis using a bigger database, with more successful instances, could try to develop a multi-class classification model able to discriminate directly among the three classes.

At this point, the dimension of our dataset was even smaller than before, indeed our data sample was constituted by only 157 campaigns and we used as train set 70% of the instances (110 campaigns) and as test set the remaining 30% (47 campaigns). Such limited size led to two main limitations. The first is that having considered a small number of campaigns, our results are true for our observations, but should be validated and generalized on a larger dataset. The second one, is that to avoid the overfitting problem (*see paragraph 6.2.2.*) we could not use all the 24 predictors (20 static and 4 dynamics) of our initial database, but we had to reduce the features to 14. In particular we excluded from our analysis most of the linguistic variables. So, differently from the unsuccess vs. success discrimination we did not manage to understand the relative importance of such factors on overfunding. For future researches it would be worthy to have a bigger database and take into account also how these factors influence the overfunding phenomenon.

Moreover, for the whole research, we used a database of Kickstarter campaigns launched and concluded between 2016-2017. This could be a limit because we considered a setting limited in space and time that may not fully represent the complexity of the phenomenon under question, but instead could bias the results. For this reason, future studies could focus on more recent campaigns, to see whether the same conditions and findings hold. Even the temporal window considered could be a limitation since some years have passed and the overfunding phenomenon has started to raise attention only in these latest years, so the same study considering more recent campaigns may bring to different results.

Finally, for our work we considered a specific crowdfunding context, namely the reward-based one, on a specific platform, Kickstarter. Being this the most important reward-based crowdfunding platform, we think that it can be fairly representative, however this could be a limitation of our results, since investigating a different crowdfunding platform governed by different dynamics may lead to different findings. Therefore, future studies could analyse different crowdfunding contexts, such as equity or donation based crowdfunding platforms.

7.5. Implications

In this paragraph we will present the main implications of our Dissertation, from both a theoretical and managerial point of view.

7.5.1. Implications for the theory

As pointed out in *Chapter 3,* our main research objective is to provide a comprehensive definition and operationalization of the overfunding concept, so to discriminate in a clear way between success and *"over-success"*, since the existing literature is unclear and does not concretely address the issue. Precisely we found that, in our empirical model, the overfunding phenomenon is defined for very high percentages of the ratio between pledged amount and target capital. The percentage defined as a discriminant is *pledge/target ≥ 175%.*

Compared to the findings already achieved by scholars, with our empirical results we find evidences supporting the definition provided by Mollik 2014, who state that a project is overfunded when the funding amount is considerably higher than the target, but we also managed to numerically understand the meaning of this definition, which seems very vague in the form proposed by the author. Overfunding, on the contrary, cannot just be defined as pledge exceeding the target (Frydrych et al., 2014; Ma X. et al., 2018; Li Y. et al., 2020; Cordova et al., 2015), as it is a too generic definition, that gives no insight on how much the pledge has to overcome the target to be in an overfunding domain.

In addition, our study provided a contribution to the literature by analysing the relative importance of 14 static and dynamic variables, and by making a comparison with the factors influencing success. While success is affected only by campaign-related variables, overfunding is impacted by both campaign-related and fundraisers-related factors, underlining how this phenomenon is conditioned also by some personal traits of the fundraiser. This result suggests that when evaluating a project investment, potential backers seem to be more attracted only by the campaign-related attributes when the same campaign has not already reached the target, while, when the campaign has reached it and so can be yet defined as successful, fundraisers-related aspects are crucial to convince backers to continue investing and to impact possible overfunded outcomes.

Another contribution to the extant literature is given by our third research objective. Indeed, while there are plenty of papers investigating the impact of the time variable on the probability of success, discovering that the first week is the most crucial to achieve a successful outcome, to the best of our knowledge, there are no papers investigating the impact of *"time"* on overfunding. Our results show that, to have a complete overview of all the overfunding campaigns, and achieve the highest accuracy of the classification model, the best option is to consider 31 days, providing an important insight in the time span that has to be considered in the overfunding case. Moreover, another contribution was provided, since we made a parallelism between success and overfunding, discovering that the first week could be used in predictive models that aim at comparing the two phenomena, with a more than good confidence level.

A final contribution to the literature is given by the application of machine learning techniques to the study of the overfunding environment. When we analysed and selected all the papers available in the literature in line with our theme (both articles, conference proceedings and books' chapters), we realised that the phenomenon has always been studied adopting statistical and qualitative reasoning and techniques. Our study with the implementation of an empirical model, through the use of MATLAB functions and machine learning algorithms, is among the first ones using this method to predict the crowdfunding campaigns' *"over-success"*.

In the end, we thoroughly believe that our thesis has opened the door to new studies and researches, in order to investigate other interesting insights about overfunding, deepening the theme with new results, using machine learning algorithms. After analysing its dynamics, we believe that overfunding is very interesting and peculiar, worthy of being widely considered among the different areas of interest related to the crowdfunding world.

7.5.2. Implications for practitioners

The practical implications of our work reach two classes of stakeholders: the crowdfunding platform involved and the project creators.

As emerged deepening all the areas explored by scholars discussed in the *Literature Review Chapter*, overfunding could lead to some positive effects but it also has several drawbacks.

As previously pointed out, the phenomenon can cause information asymmetries leading to the *"market for lemon"* environment and, therefore, to the exit of high-quality projects, impeding the ultimate goal of crowdfunding, hindering open innovation. Many practitioners, such as Koch J. and colleagues (2018), state that massively overfunded projects could overshadow other campaigns which, in turn, receive less money, and try to *"modify"* the crowdfunding platform, proposing a taxation mechanism to internalize these overfunding negative externalities thus improving the overall funding results. By applying our predictive tool, a similar objective can be achieved. Indeed, a crowdfunding platform can be modified, in order to offer a service to the fundraisers of the campaigns: our predictive model opens the possibility to monitor the trend of a campaign day by day also providing the *"predictive"* outcome, namely visualize with a certain confidence level the expected result of the campaign. A section dedicated to the fundraiser could be created in the platform, with a complete part displaying all the quantitative and qualitative project KPIs and factors. In the latter section it would be possible to see, since the campaign launch, whether the project will reach success or even be overfunded, given certain parameters and with a certain level of accuracy. As the days progress, the section will be updated considering also dynamic factors.

Regarding the fundraisers, they should be aware that, while a successful outcome is highly desired, an overfunded outcome could be not. As emerged in the papers provided by scholars, overfunding can be detrimental for overly successful fundraisers because the cash surplus is unlikely to be effectively managed, as it has not been budgeted in the original business plans (M.A. Stanko et al., 2017). For example, taking into consideration a reward-based crowdfunding platform, such as Kickstarter, a higher level of overfunding is related to the higher costs of the extra rewards that should be delivered to additional project backers (Makýšová L. et al., 2017). In addition, taking a psychological perspective, overfunding could also pose a threat to overly successful fundraisers by inflating their egos and spurring them to take on unnecessary risks. As a consequence, overfunded campaigns tend to become overly ambitious, culminating in unrealistic claims, unfulfilled obligations, and ultimately, disappointed investors (Li Y. et al., 2020). For this reason, firstly fundraisers should take into considerations the factors affecting overfunding and strictly monitor them, to be sure that they achieve a manageable amount of money. The intrinsic value of the project launched (target and project category) remains still an important factor in their evaluation, but, in addition, fundraisers should also carefully consider their personal profile and information, since, as demonstrated by the results obtained, differently from the prediction of success, for the overfunding, also fundraisers-related factors are relevant.

7.6. Conclusions

The study of overfunding, a peculiar phenomenon that influences crowdfunding campaigns is still in an early stage, and, after having reviewed the extant literature, we can affirm that there is not a unanimous result yet. Indeed, there are grounds for broadening knowledge in this direction. In our analysis we employed machine learning models as a tool to predict the outcome of a campaign starting from the data available to the backers. This methodology can be seen as a novel approach to address this interrogative.

The use of these models led us to identify firstly a precise definition and operationalisation of overfunding, embracing the definition provided by Mollick (2014): *"Overfunding is especially used when a project's funding is considerably higher than its funding goal",* but also managing to concretely define the general concept of *"considerably higher"* into mathematical terms. Indeed, the development of a classification empirical model, highlighted that, through the definition of a high-overfunding threshold, it is possible to discriminate between the two phenomena of success and overfunding with a certain confidence level related to the accuracy of the model provided.

It is noteworthy that machine learning algorithms have an important advantage compared to other techniques employed in precedent studies (e.g., statistical analysis). Indeed, we managed to investigate, relying on two different techniques of features' ranking, the impact of all the static and dynamic factors on the campaigns' *"over-success"*, also ranking them based on their relative importance for the predicting models. The result of this process revealed that, among static predictors, campaign-related factors (*s_target* and *s_category*) and fundraisers-related factors (*s_seria*l and *s_team*) are the most important in determining the outcome of the campaigns. At the same time, among the dynamic predictors, the most relevant one is *d_percentpledge,* underlining that the accuracy of the classification tool drastically increases as the days of the campaign progress, overcoming a level of 98,20% in day 31st, implying that these dynamic variables have a great positive impact on the capability of the model to predict the overfunding of a campaign.

Moreover, through our study we managed to understand the most important time horizon to be considered in the overfunding case, also providing a parallelism with the findings related to the success domain. In order to have a complete overview of the phenomenon and reach the best outcome possible, in terms of accuracy of the classification model, the entire life-span of a project should be considered. However, with a good degree of approximation the first seven days can be seen as relevant in predicting the probability of overfunding, as in the case of success.

Our work has practical implications for fundraisers, as they should focus on factors affecting overfunding, bearing in mind that while a successful outcome is highly desired, an overfunded outcome could be not, since it may bring several negative implications for them.

In conclusion, our contributions have some limitations, and these open up opportunities for further research. In particular, it would be interesting to generalize our findings investigating how the accuracy of the predictive tool changes considering a bigger sample to train the models. Furthermore, in order to assess whether the relative importance of the success factors has changed overtime, it would be very insightful to consider more recent campaigns. Finally, an additional research opportunity is given by the possibility of replying this work on contexts different form the reward-based one, since different crowdfunding environments are characterised by specific dynamics, and this could affect both the overfunding definition and the relative importance of the predictors.

REFERENCES

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-Based Learning Algorithms. Machine Learning, 6, 37–66. https://doi.org/10.1023/A:1022689900470.

Ahlers, G. K. C., Cumming, D., Günther, C., & Schweizer, D. (2015). Signalling in Equity Crowdfunding. *Entrepreneurship: Theory and Practice*, *39*(4), 955–980. https://doi.org/10.1111/etap.12157

Andrei Hagiu and Julian Wright (2015). Multi-Sided-Platforms. *International Journal of Industrial Organization*, 43, 162-174.

Annunziata and Aversa (2020). Predicting the success of crowdfunding campaigns: a machine learning approach. *Politecnico di Milano.*

Barbi, Massimiliano & Bigelli, Marco, 2017. Crowdfunding practices in and outside the US, *Research in International Business and Finance*, Elsevier, vol. 42(C), pages 208-223.

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2010). Crowdfunding: An industrial organization perspective. *Prepared for the Workshop Digital Business Models: Understanding Strategies', Held in Paris on June. 2010.*, 57–64.

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, *29*(5), 585–609.

https://doi.org/https://doi.org/10.1016/j.jbusvent.2013.07.003

Bishop C. (2006). Pattern Recognition and Machine Learning. *Springer.*

Block, J., Hornuf, L., & Moritz, A. (2018). Which updates during an equity crowdfunding campaign increase crowd participation? *Small Business Economics*, *50*(1), 3–27. https://doi.org/10.1007/s11187-017-9876-4

Boeuf, B., Darveau, J., & Legoux, R. (2014). Financing creativity: Crowdfunding as a new approach for theatre projects. *International Journal of Arts Management*, *16*(3).

Burkett, E. (2011). A Crowdfunding Exemption? Online Investment Crowdfunding and U.S. Securities Regulation. *Transactions: The Tennessee Journal of Business Law*, *13*, 63.

Burtch, G., Ghose, A., & Wattal, S. (2013). An empirical examination of users' information hiding in a crowdfunding context.

Butticé, V., Franzoni, C., Rossi-Lamastra, C., & Rovelli, P. (2018). The road to crowdfunding success: A review of the extant literature. In *Creating and Capturing Value through Crowdsourcing*. https://doi.org/10.1093/oso/9780198816225.003.0005.

Buttice, V., & Useche, D. (2019). Crowdfunding to Overcome Liability of Outsidership: Drivers of Immigrants' Fundraising Performance. *Academy of Management Proceedings*. https://doi.org/10.5465/ambpp.2019.11642abstract.

Cai, W., Polzin, F., & Stam, E. (2020). Crowdfunding and social capital: A systematic review using a dynamic perspective. *Technological Forecasting and Social Change*, *162*. https://doi.org/10.1016/j.techfore.2020.120412.

Calic, G., & Mosakowski, E. (2016). Kicking Off Social Entrepreneurship: How A Sustainability Orientation Influences Crowdfunding Success. *Journal of Management Studies*, *53*(5), 738–767. https://doi.org/10.1111/joms.12201.

Cicchiello A.F., Kazemikhasragh A., Monferrà S. (2020). Gender differences in new venture financing: evidence from equity crowdfunding in Latin America. *International Journal of Emerging Markets.*

Chan, C. S. R., & Parhankangas, A. (2017). Crowdfunding Innovative Ideas: How Incremental and Radical Innovativeness Influence Funding Outcomes. *Entrepreneurship: Theory and Practice*, *41*(2), 237–263.

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, 12(4), 50-(4), 50–50.

Chen, S., Thomas, S., & Kohli, C. (2016). What Really Makes a Promotional Campaign Succeed on a Crowdfunding Platform? *Journal of Advertising Research*, *56*(1), 81–94. https://doi.org/10.2501/jar-2016-002.

Chen, Y., Zhang, W., Yan, X., & Jin, J. (2020). The life-cycle influence mechanism of the determinants of financing performance: an empirical study of a Chinese crowdfunding platform. *Review of Managerial Science*, *14*(1), 287-309.

Colombo, M. G., Franzoni, C., & Rossi-Lamastra, C. (2015a). ENTREPRENEURSHIP. Cash from the crowd. *Science (New York, N.Y.).*

Colombo, M. G., Franzoni, C., & Rossi-Lamastra, C. (2015b). Internal social capital and the attraction of early contributions in crowdfunding. *Entrepreneurship: Theory and Practice*, *39*(1), 75–100. https://doi.org/10.1111/etap.12118.

Alessandro Cordovaa , Johanna Dolcib , Gianfranco Gianfratec (2015), The determinants of crowdfunding success: evidence from technology projects, *3rd International Conference on Leadership, Technology and Innovation Management.*

Crosetto, P., & Regner, T. (2014). Crowdfunding: Determinants of success and funding dynamics. *Jena Economic Research Papers No. 2014-035*.

Davidsson, P., & Honig, B. (2003). The role of social and human capital among nascent entrepreneurs. *Journal of Business Venturing*, *18*(3), 301–331.

https://doi.org/10.1016/S0883-9026(02)00097-6

Doshi, A. (2014). "The Impact of High Performance Outliers on Two-Sided Platforms: Evidence fromCrowdfunding." In:Working Paper(Version: 2014/10/20).

Du S., Peng J., Nie T., Yu Y. (2020), Pricing strategies and mechanism choice in reward-based crowdfunding, *European Journal of Operational Research.*

Dutton, D. M., & Conroy, G. V. (1997). A review of machine learning. *Knowledge Engineering Review*, 12(4), 341–367. https://doi.org/10.1017/S026988899700101X

Frydrych, D., Bock, A. J., Kinder, T., & Koeck, B. (2014). Exploring entrepreneurial legitimacy in reward-based crowdfunding. *Venture Capital*, *16*(3), 247–269. https://doi.org/10.1080/13691066.2014.916512

Garry Bruton, Susanna Khavul, Donald Siegel, Mike Wright (2015/1). New financial alternatives in seeding entrepreneurship: Microfinance, crowdfunding, and peer–to–peer innovations. *SAGE Publications*. 9-26.

Gerber, E. M., & Hui, J. (2013). Crowdfunding: Motivations and deterrents for participation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *20*(6), 1-32.

Gershenson, C. (2003). Artificial Neural Networks for Beginners.

Gleasure, R., & Feller, J. (2014). Observations of Non-linear Information Consumption in Crowdfunding. In *Mining Intelligence and Knowledge Exploration* (pp. 372–381). *Springer,* Cham. https://doi.org/10.1007/978-3-319-13817-6_36

Glenn G. (2021). Training new researchers on how to understand and develop their profiles.: *University of Nebraska Medical Center. https://www.elsevier.com/__data/assets/pdf_file/0006/1157748/Scopus_UNMC-Teaching-and-Learning_CS-SC-FINAL-WEB.pdf.*

Go-Globe Startups (2017), Startups Success and Failure Rate – Statistics and Trends. https://www.go-globe.com/startups/

Greenberg, J., & Mollick, E. (2017). Activist Choice Homophily and the Crowdfunding of Female Founders*. *Administrative Science Quarterly*, *62*(2), 341–374. https://doi.org/10.1177/0001839216678847

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*, 986–996. https://doi.org/10.1007/978-3-540-39964-3_62.

HannahForbes, DirkSchaefer (2017). Guidelines for Successful Crowdfunding. *Procedia CIRP*, Volume 60, 2017, Pages 398-403

Hobbs, J., Grigore, G., & Molesworth, M. (2016). Success in the management of crowdfunding projects in the creative industries. *Internet Research*.

https://doi.org/10.1108/IntR-08-2014-0202

Jerry Coakley, Aristogenis Lazos and Jose Liñares-Zegarra (2018), Follow-onequity crowdfunding. *Business and Local Government Data Research Centre,Essex Finance Centre, and Essex Business School.*

Kaartemo, V. (2017). The Elements of a Successful Crowdfunding Campaign: A Systematic Literature Review of Crowdfunding Performance. *International Review of Entrepreneurship*, *15*(3), 291–318.

Kim, T., Por, M.H., & Yang, S.B. (2017). Winning the crowd in online fundraising platforms: The roles of founder and project features. *Electronic Commerce Research and Applications*, 25, 86-94.

K. Kim, S. Viswanathan, The Experts in the Crowd: The Role of Reputable Investors in a Crowdfunding Market, in: *The 41st Research Conference on Communication, Information and Internet Policy*, 2014.

Kučerová Z., Dařena F., 2019, Behavioural Insights from Crowdfunding Financing: Power of Nudges, *Conference on International Finance 2019.*

Koch J. (2016). The phenomenon of project overfunding on online crowdfunding platforms - Analyzing the drivers of overfunding. *24th European Conference on Information Systems, ECIS 2016*

Koch, J.-A., & Siering, M. (2019). The recipe of successful crowdfunding campaigns. *Electronic Markets*, *29*(4), 661–679. https://doi.org/10.1007/s12525-019-00357-8.

Koch J.-A., Lausen J., Kohlhase M. (2018). Towards internalizing the externalities of overfunding - Introducing a 'tax' on crowdfunding platforms. *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018.*

Koch, J.-A., & Cheng, Q. (2016). The role of qualitative success factors in the analysis of crowdfunding success: evidence from Kickstarter. *Banking & Insurance EJournal*.

Kotsiantis S.B., I. D. Zaharakis , P. E. Pintelas (2007). Machine learning: a review of classification and combining techniques. *Springer Science Business Media B.V.* 2007.

Kunz, M. M., Bretschneider, U., Erler, M., & Leimeister, J. M. (2017). An empirical investigation of signaling in reward-based crowdfunding. In *Electronic Commerce Research*, 17(3). *Springer* US. https://doi.org/10.1007/s10660-016-9249-0.

Kuppuswamy, V., & Bayus, B. L. (2017). Does my contribution to your crowdfunding project matter? *Journal of Business Venturing*, *32*(1), 72–89. https://doi.org/10.1016/j.jbusvent.2016.10.004.

Lagazio, C., & Querci, F. (2018). Exploring the multi-sided nature of crowdfunding campaign success. *Journal of Business Research*, *90*(May), 318–324.

https://doi.org/10.1016/j.jbusres.2018.05.031

Lasrado, L. A., & Lugmayr, A. (2013). Crowdfunding in Finland: A New Alternative Disruptive Funding Instrument for Businesses. *Proceedings of International Conference on Making Sense of Converging Media*, 194–201.

https://doi.org/10.1145/2523429.2523490

Li Y., Liu F., Fan W., Lim E.T.K., Liu Y. (2020). Exploring the impact of initial herd on overfunding in equity crowdfunding. *Information and Management*

Liao, C., Zhu, Y., & Liao, X. (2015). The role of internal and external social capital in crowdfunding: Evidence from China. *Revista de Cercetare Si Interventie Sociala*, *49*, 187–204.

Lindgren, T. (2004). Methods for rule conflict resolution. European Conference on Machine Learning, 262–273. https://doi.org/10.1007/978-3-540-30115-8_26

Liu, J., L. Yang, Z. Wang and J. Hahn (2015). Winner Takes All? The "Blockbuster Effect" inCrowdfunding Platforms. *Proceedings of the Thirty Sixth International Conference on Infor-mation Systems (*ICIS'15).

Ma X., Yang M., Li Y., Zhang J. (2017). Signaling factors in overfunding: An empirical study based on Crowdcube. *14th International Conference on Services Systems and Services Management, ICSSSM 2017 – Proceedings.*

Makýšová L., Vaceková G. (2017). Profitable Nonprofits? Reward-Based Crowdfunding in the Czech Republic. *NISPAcee Journal of Public Administration and Policy.*

Malave, I. (2012). Why Kickstarter Should More Fully Integrate Social Media. In:*Working Paper(*Version: 2012/11).Mollick, E. (2014).

Martens, M. L., Jennings, J., & Jennings, P. (2007). Do the stories they tell get them the money they need? The role of entrepreneurial narratives in resource acquisition. *Academy of Management Journal*, *50*(5), 1107–1132.

Martínez-Gómez C., Jiménez-Jiménez F., Alba-Fernández M.V. (2020). Determinants of overfunding in equity crowdfunding: An empirical study in the UK and Spain. *Sustainability (Switzerland).*

Miro Arola (2018). Campaign overfunding and bounded rationality in equity crowdfunding, *Aalto Business School.*

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. *33rd Annual Frontiers in Education - Frontiers in Education Conference*, FIE, T2A-13. https://doi.org/10.1109/FIE.2003.1263284.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, https://doi.org/10.1016/j.jbusvent.2013.06.005.

Mollick, E., & Nanda, R. (2016). Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science*, *62*(6), 1533–1553. https://doi.org/10.1287/mnsc.2015.2207.

Moritz A., J. Block, E. Lutz, Investor communication in equity-based crowdfunding: a qualitative-empirical study*, Qual. Res. Financ. Mark.* 7 (2015) 309–342.

Muggleton, S. (1999). Inductive Logic Programming: Issues, results and the challenge of Learning Language in Logic. *Artificial Intelligence*, 114(1–2), 283–296.

https://doi.org/10.1016/s0004-3702(99)00067-3

Nielsen, K. R., & Binder, J. K. (2020). I Am What I Pledge: The Importance of Value Alignment for Mobilizing Backers in Reward-Based Crowdfunding. *Entrepreneurship: Theory and Practice*.

https://doi.org/10.1177/1042258720929888

Omary Z., Fredrick Mtenzi (2010), Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning. *International Journal for Infonomics (IJI), Volume 3, Issue 3, September 2010.*

Ordanini, A., & Parasuraman, A. (2011). Service innovation viewed through a service-dominant logic lens: A conceptual framework and empirical analysis. *Journal of Service Research*, *14*(1), 3–23. https://doi.org/10.1177/1094670510385332.

Patil, D. D., Wadhai, V. M., & Gokhale, J. A. (2010). Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy. *International Journal of Computer Applications,* 11(2), 23–30. https://doi.org/10.5120/1554-2074.

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology, 49(5), 871–887.* https://doi.org/10.1002/ejsp.2561.

Pietraszkiewicz, A., Soppe, B., & Formanowicz, M. (2017). Go Pro Bono Prosocial Language as a Success Factor in Crowdfunding. *Social Psychology*, 48(5), 265–278. https://doi.org/10.1027/1864-9335/a000319.

Pitschner, S., & Pitschner-Finn, S. (2014). Non-profit differentials in crowd-based financing: Evidence from 50,000 campaigns. *Economics Letters*, *123*(3), 391–394. https://doi.org/10.1016/j.econlet.2014.03.022.

Quinlan, J. R. (1993). C4.5: programs for machine learning Morgan Kaufmann *Publishers Inc., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.*

Raafat R.M., N. Chater, C. Frith, Herding in humans, *Trends Cogn. Sci*. (Regul. Ed.) 13 (2009) 420–428.

Radzicki M. J. System Dynamics and Its Contribution to Economics and Economic Modeling. *Department of Social Science and Policy Studies, Worcester Polytechnic Institute, Worcester, MA, USA*.

Riedl, J. (2013). Crowdfunding technology innovation. *Computer*, (3), 100-103.

Rokach, L., & Maimon, O. (2006). Decision Trees. In Data Mining and Knowledge Discovery Handbook (pp. 165–192). *Springer*. https://doi.org/10.1007/0-387-25465-x_9.

Ryoba, M. J., Qu, S., & Zhou, Y. (2020). Feature subset selection for predicting the success of crowdfunding project campaigns. *Electronic Markets*, 1–14. https://doi.org/10.1007/s12525-020-00398-4.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386. https://doi.org/10.1037/h0042519.

Sá, J. A. S., Almeida, A. C., Rocha, B. R. P., Mota, M. A. S., Souza, J. R. S., & Dentel, L. M. (2016). Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy. https://doi.org/10.21528/cbic2011-27.1.

Samuel, A. L. (1959). Some Studies in Machine Learning. *IBM Journal of Research and Development.*

Schapire Rob, Machine Learning Algorithms, *Princeton University*.

Scholkopf, B., Burges, C. J. C., & Smola, A. J. (1998). Advances in Kernel Methods - Support Vector Learning. *MIT Press*.

Schwienbacher, A., & Larralde, B. (2010). Crowdfunding of small entrepreneurial ventures. *Handbook of entrepreneurial finance, Oxford University Press, Forthcoming.*

Skirnevskiy, V., Bendig, D., & Brettel, M. (2017). The influence of internal social capital on serial creators' success in crowdfunding. *Entrepreneurship Theory and Practice*, *41*(2), 209-236.

Stanko M.A., D.H. Henard, Toward a better understanding of crowdfunding, openness and the consequences for innovation, *Res. Policy* 46 (2017) 784–798.

Supriya, M., & Deepa, A. J. (2020). Machine learning approach on healthcare big data: a review. *Big Data and Information Analytics*, 5(1), 58–75. https://doi.org/10.3934/bdia.2020005.

Svatopluk Kapounek, Zuzana Kučerová (2019). Overfunding and Signaling Effects of Herding Behavior in Crowdfunding. *CESifo Working Paper No. 7973.*

Svidronová M., Vaceková, G M. Plaček, Markéta Matulová, Lucia Hrůzová, Lenka Harringová (2020), Alternative non-profit funding methods: crowdfunding in the Czech Republic and Slovakia. *Applied Economics Letters.*

Theerthaana and Manzoor, 2019, A signalling paradigm incorporating an Agent-Based Model for simulating the adoption of crowd funding technology, *Journal of simulation, computer science.*

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, 45(3), 251–269. https://doi.org/10.1023/B:RIHE.0000019589.79439.6e.

Turan S.S., Stakeholders in equity-based crowdfunding: respective risks over the equity crowdfunding lifecycle*, J. Financ. Innov.* 1 (2015) 141–151.

Vercellis, C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. *Business Intelligence: Data Mining and Optimization for Decision Making. https://doi.org/10.1002/9780470753866.*

Wang, W., Zhu, K., Wang, H., & Wu, Y. C. J. (2017). The impact of sentiment orientations on successful crowdfunding campaigns through text analytics. *IET Software*, *11*(5), 229–238. https://doi.org/10.1049/iet-sen.2016.0295.

Witten I. H. and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques.

Yen C.-H., Lee Y.-C., Fu W.-T. (2018). Visible hearts, visible hands: A Smart Crowd donation platform. *International Conference on Intelligent User Interfaces, Proceedings IUI.*

Yuan, X., & Wang, H. (2020). How linguistic cues affect the motivation of capital-giving in crowdfunding: a self-determination theory perspective. *New Review of Hypermedia and Multimedia*. https://doi.org/10.1080/13614568.2020.1802518.

Zheng, H., Li, D., Wu, J., & Xu, Y. (2014). The role of multidimensional social capital in crowdfunding: A comparative study in China and US. *Information and Management*, *51*(4), 488–496. https://doi.org/10.1016/j.im.2014.03.003.

Zvilichovsky, D., Inbar, Y., & Barzilay, O. (2013). Playing both sides of the market: Success and reciprocity on crowdfunding platforms. *SSRN Electronic Journal.*, *4*. https://doi.org/10.2139/ssrn.2304101.

ANNEX 1

Histograms of the static numeric variables:

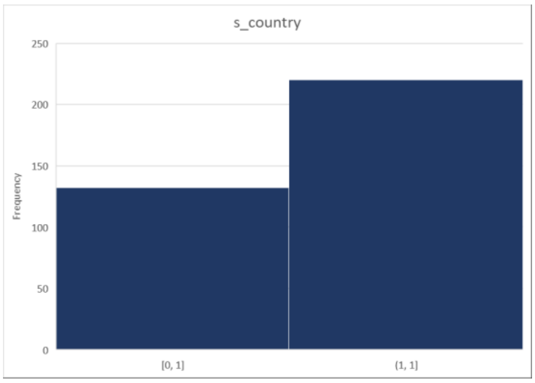


Figure 75: s_country – histogram

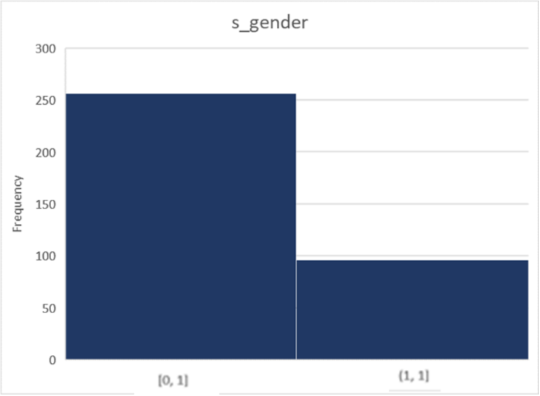


Figure 76: s_gender – histogram

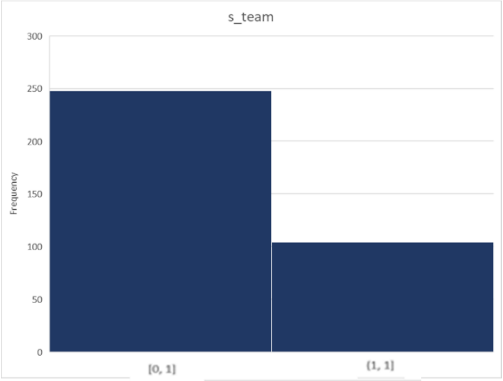


Figure 77: s_team – histogram

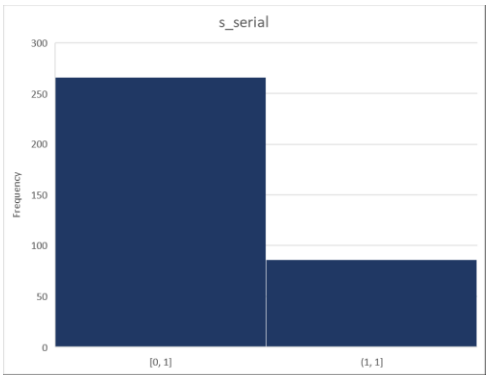


Figure 78: s_serial – histogram

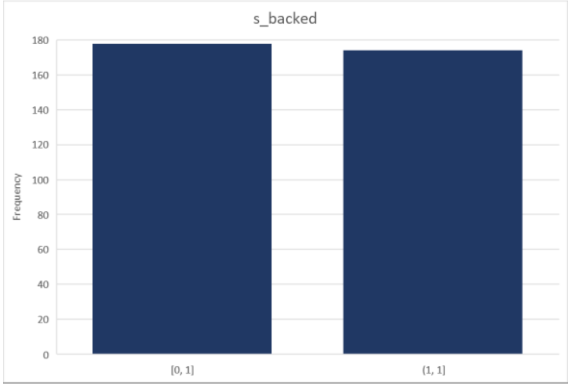


Figure 79: s_backed – histogram

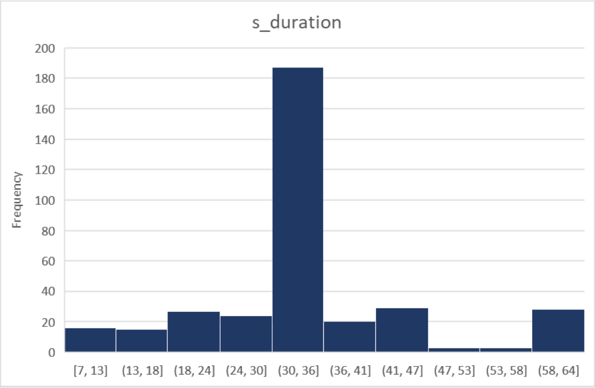


Figure 80: s_duration – histogram

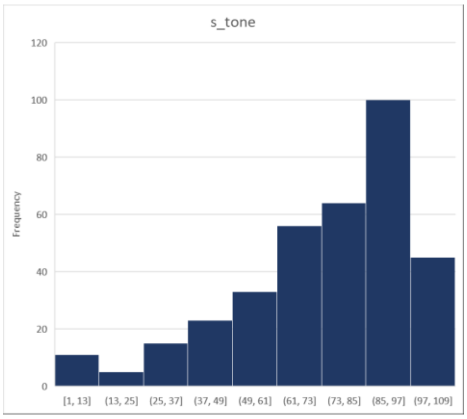


Figure 81: s_tone – histogram

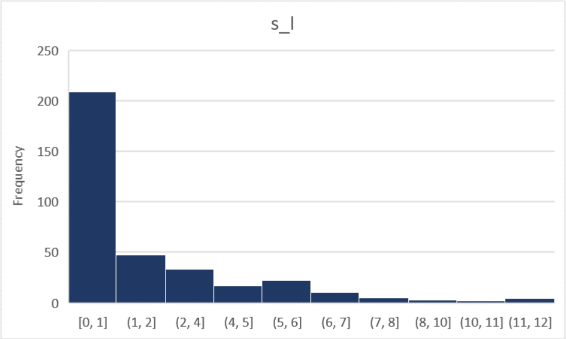


Figure 82: s_i – histogram

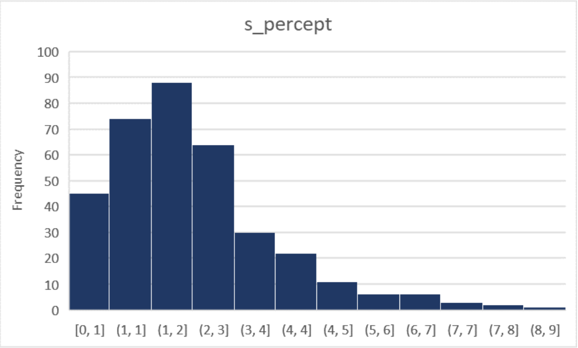


Figure 83: s_percept – histogram

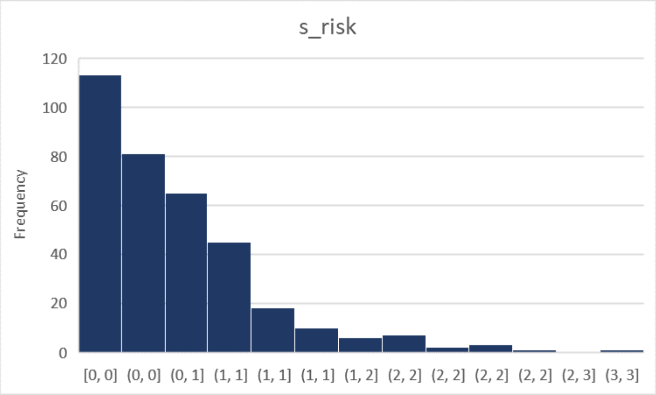


Figure 84: s_risk – histogram

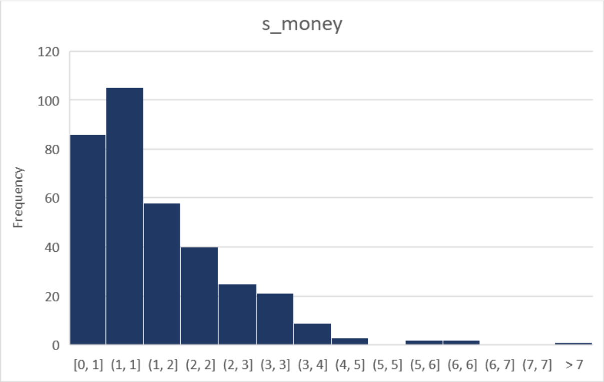


Figure 85: s_money – histogram

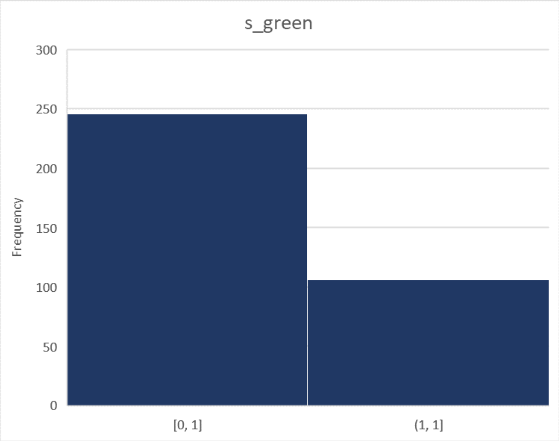


Figure 86: s_green – histogram

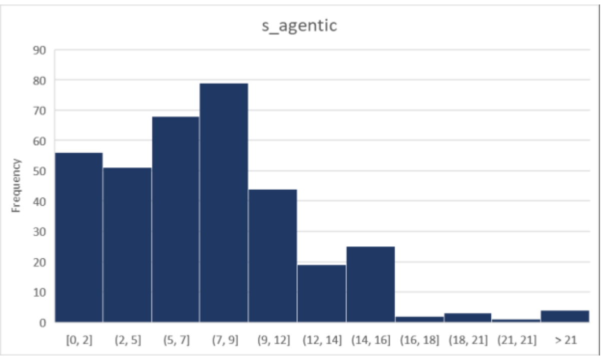


Figure 87: s_agentic – histogram

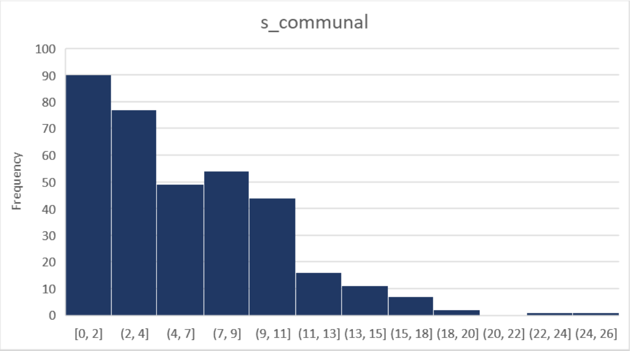


Figure 88: s_communal – histogram

Boxplots of the static numeric variables:

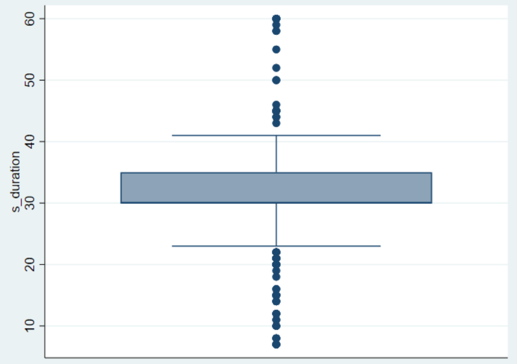


Figure 89: s_duration – boxplot

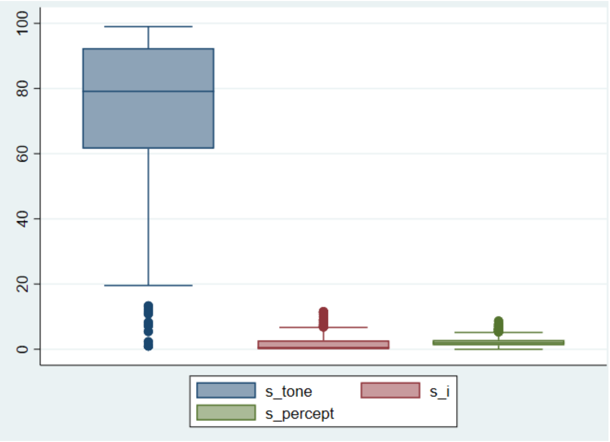


Figure 90: s_tone, s_percept, s_i – boxplot

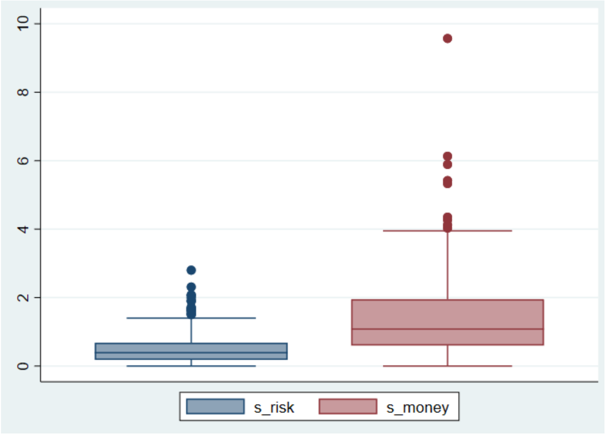


Figure 91: s_risk, s_money – boxplot

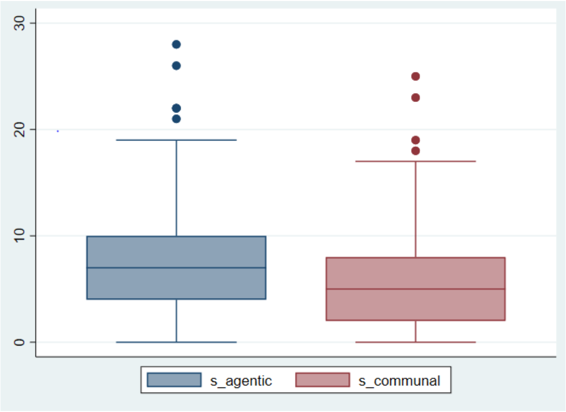


Figure 92: s_agentic, s_communal – boxplot

ANNEX 2

Trend overfunding percentage:

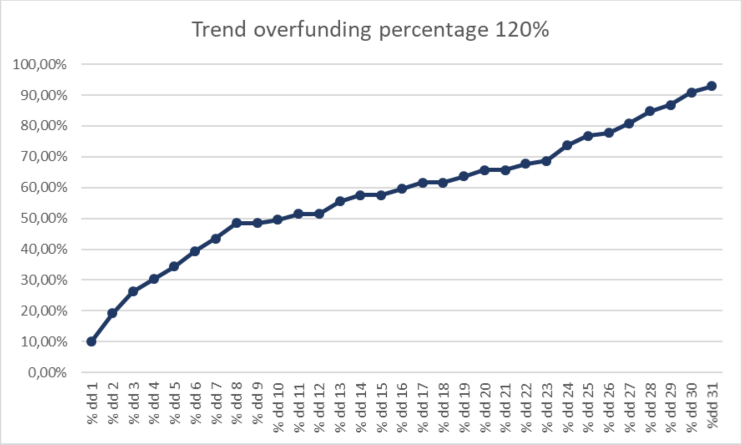


Figure 93: Trend overfunding percentage 120%

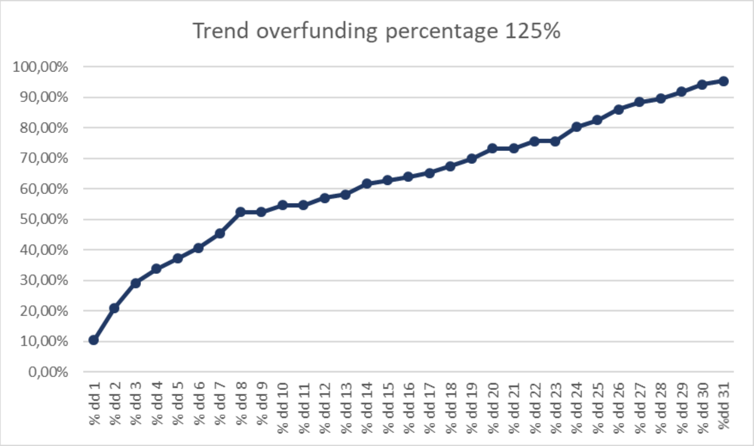


Figure 94: Trend overfunding percentage 125%

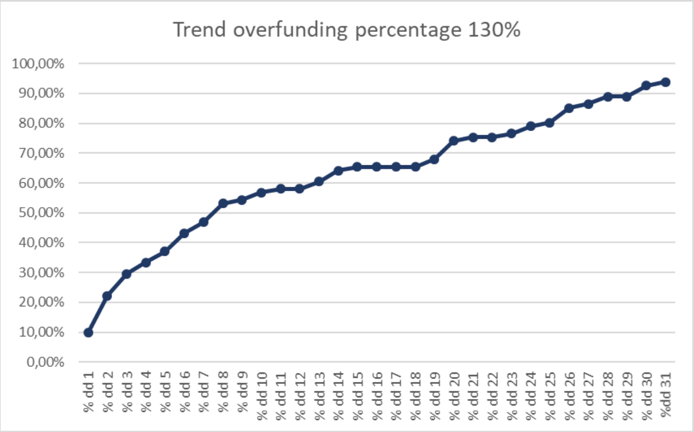


Figure 95: Trend overfunding percentage 130%

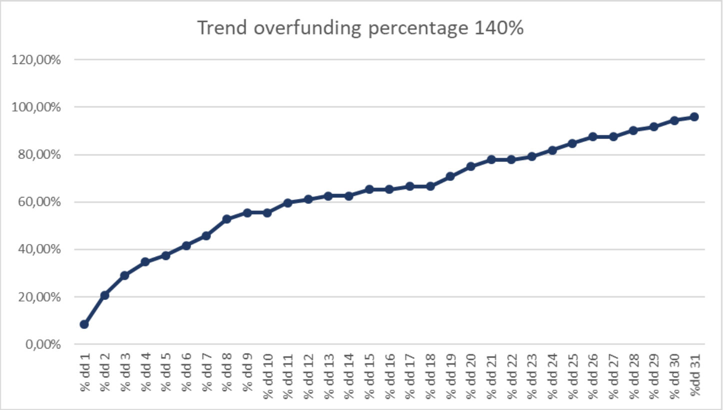


Figure 96: Trend overfunding percentage 130%

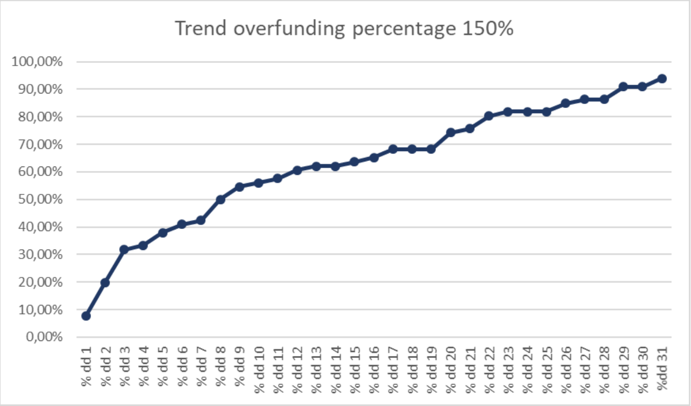


Figure 97: Trend overfunding percentage 150%

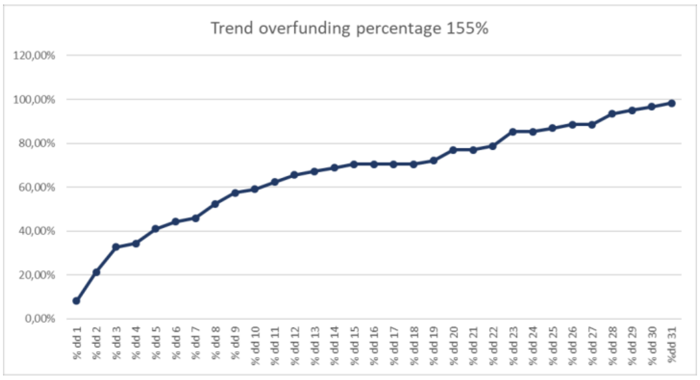


Figure 98: Trend overfunding percentage 155%

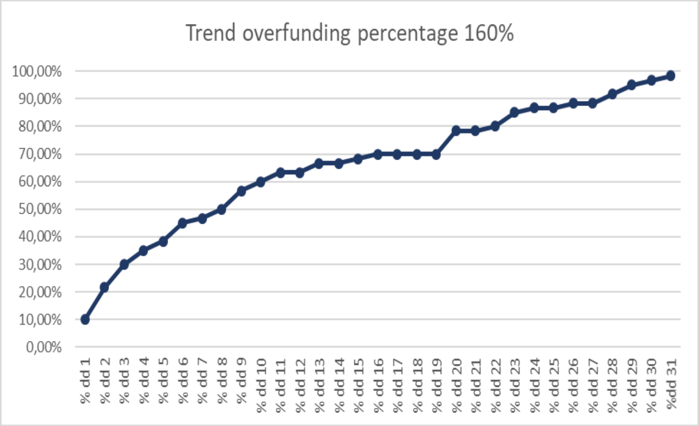


Figure 99: Trend overfunding percentage 160%

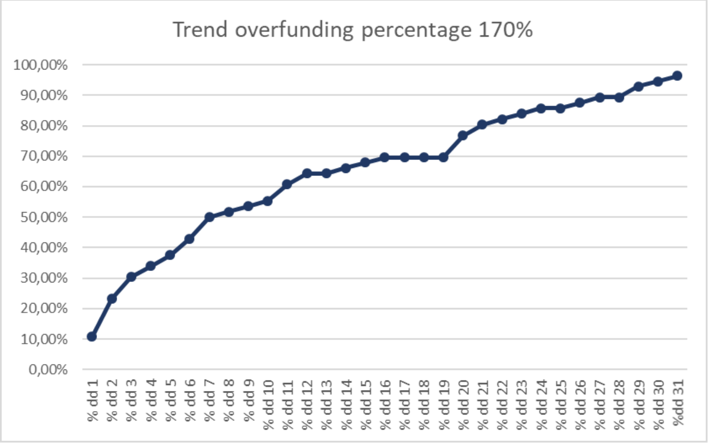


Figure 100: Trend overfunding percentage 170%