POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Customer churn prediction in a slow fashion e-commerce context: an analysis of the effect of static data in customer churn prediction

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Author: **Luca Colasanti**

Student ID: 968221
Advisor: Prof. Marcello Restelli
Co-advisors:
Academic Year: 2021-22

Abstract

Survival analysis is a subfield of statistics where the goal is to analyse and model the data where the outcome is the time until the occurrence of an event of interest. Because of the intrinsic temporal nature of the analysis, the employment of more recently developed sequential models (Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM)) has been paired with the use of dynamic temporal features, in contrast with the past reliance on static ones. Such an abrupt shift of policy has left open the challenge of understanding how those two kinds of features influence the predictive capabilities of models. This thesis aims at assessing the effect of combining static and dynamic features on the most commonly used models in survival analysis. In doing so, we compare the error measurements of such models with dataset composed of purely dynamic features or a combination of static and dynamic ones. Empirical measurements have shown that models respond differently to the addition of static features to the analysis, with more complex, sequential models like the LSTM struggling to deal with the added data complexity (with a 12% increase in error), while non sequential models see reductions of up to 14.7% in error. The thesis also includes a clusterization task aimed at aiding the interpretation of survival analysis outcomes.

Keywords: Survival Analysis, Time To Event prediction, Churn retention, Machine Learning, Deep Learning, Customer Clustering, E-commerce

Abstract in lingua italiana

L'analisi della sopravvivenza è una branca della statistica il cui obiettivo è l'analisi e la modellazione di dati il cui risultato è il tempo che intercorre fino al verificarsi di un evento di interesse. A causa dell'intrinseca natura temporale dell'analisi, l'impiego di modelli sequenziali di più recente sviluppo (RNN e LSTM) è stato abbinato all'uso di attributi temporali dinamici, a differenza dell'uso più diffuso in passato di attributi statici. Questo brusco cambiamento ha lasciato aperta la sfida di capire come questi due tipi di attributi influenzino le capacità predittive dei modelli. Questa tesi si propone di valutare l'effetto della combinazione di attributi statici e dinamici sui modelli più comunemente utilizzati nell'analisi della sopravvivenza. A tal fine, confrontiamo le misure di errore di tali modelli con set di dati composti da attributi puramente dinamici o da una combinazione di statici e dinamici. I risultati empirici hanno mostrato che i modelli rispondono in modo diverso all'aggiunta di attributi statici, con i modelli sequenziali più complessi, come l'LSTM, che faticano a gestire la complessità dei dati aggiunti (con un aumento dell'errore del 12%), mentre i modelli non sequenziali registrano riduzioni dell'errore fino al 14,7%. La tesi comprende anche una clusterizzazione volta a facilitare l'interpretazione dei risultati dell'analisi di sopravvivenza.

Parole chiave: Analisi di sopravvivenza, Previsione del tempo a evento, Ritenzione dall'abbandono dei clienti, Apprendimento automatico, Apprendimento profondo, Segmentazione della clientela, Commercio elettronico

Acronyms

CDF Cumulative Distribution Function
E-commerce Electronic Commerce
SME Small and Medium Enterprises
SMB Small and Medium Businesses
CAC Customer Acquisition Cost
ML Machine Learning
RFM Recency-Frequency-Monetary
CLV Customer Lifetime Value
DL Deep Learning
ANN Artificial Neural Network
SOM Self-Organising Map
MADM Multi Attribute Decision Making
ECPR Employee Churn Prediction and Retention
TOPSIS Technique for Order Preference by Similarity to Ideal Solution
DT Decision Tree
LR Logistic Regression
LLM Logit Leaf Model
SVM Support Vector Machine
TTE Time To Event
WTTE-RNN Weibull Time To Event Recurrent Neural Network
RUL Remaining Useful Life
RL Reinforcement Learning
RNN Recurrent Neural Network
LSTM Long Short Term Memory
CDF Cumulative Distribution Function
KM Kaplan-Mayer approach
CRM Customer Relationship Management
MAE Mean Average Error
MSE Mean Squared Error

MLP	Multi-Layer Perceptron
PCA	Principal Component Analysis
FAMD	Factory Analysis of Mixed Data
SEK	Swedish Krona
TOE	Time Of Event
ET	Execution Time
RF	Random Forest
MCA	Multiple Correspondence Analysis
AI	Artificial Intelligence
DNN	Deep Neural Network
KNN	K Nearest Neighbours

Contents

1 | Introduction

In this thesis a case study on the effects of static and dynamic features on a survival analysis task will be presented. The context of analysis is that of a slow fashion e-commerce and, in more detail, the thesis will revolve around the prediction of a specific event, customer churn, assessed through multi model learning.

1.1. Background

As defined by [77], a customer is labelled as a churner if they have no events (purchases in this instance) in a set period of time. Thus, the "customer churn" event is the point in time after which a customer has no purchases. In the specifics of a fashion application, churn rate becomes central in ensuring the profitability and survivability of a business: as stated by [72] already in 2006, poor customer management practices and non-personalized marketing efforts result in customers becoming dissatisfied with a company, churning and buying from a competitor. The consequence of sustained loss of customers is the reason why companies are forced to increase their efforts in customer acquisition processes, but the growing competition in the Electronic Commerce (E-commerce) environment is making customer acquisition processes and customer retention processes much more expensive than before, with industry stalwarts seeing their Customer Acquisition Cost (CAC) up 70 to 75% whereas new markets are seeing increases closer to 50% over the past five years [74]. It's shown that it costs five times as much to attract a new customer than to keep an existing one [49] and that a customer who makes multiple purchases spends on average four times as much money as a customer who makes only one purchase [3]. In order to mediate the increase in costs, the most relevant solution is to focus and invest more in customer retention processes.
Another aspect relates to the emergence of slow fashion, which emphasizes slowing down both the production and the consumption processes, encouraging sustainable values among all who partake in the fashion system [20][61]. [53] provides a theoretical definition of slow fashion in 5 points, one of which is: maximizing product lifespan and efficiency for a sustainable environment (functionality). This is in contrast with fast fashion, shaped by a set

of business practices focused on achieving continual economic growth [30], often obtained by pushing new trends and aiming at mass consumption. Due to the fashion industry being considered one of the most polluting industries in the world [17], customers' influence on business decisions [40] has led companies to open up to more sustainable production and sale models, which can benefit from better customer retention practices.

The problem of churn retention represents an instance of survival analysis, a subfield of statistics where the goal is to analyse and model the data where the outcome is the time until the occurrence of an event of interest [95]. Survival analysis was developed to assess patient survival, and while death is often the primary event of interest, survival analysis can also be used to assess treatment failure, such as time to loss of graft patency or amputation. Rather than simply addressing frequency, survival analysis also captures an element of time to an event. It also incorporates censorship, in which data about the event of interest are unknown because of withdrawal of the patient from the study [75]. Survival analysis was first developed through different statistical models, like the Kaplan-Meyer [35] and Cox Proportional Methods [93] that are further referenced and analysed in 2.1.

Survival analysis has seen applications in many different fields over the years, from medical to business environments, but in many of the early applications in the Machine Learning (ML) field there has been a reliance on the use of static features fed to non sequential models. Works like [6], [90], [50], [45], [98], [24] either compare different ML architectures or propose new frameworks applied to different iterations of survival analysis problems, with the main aspect being the use of *static* features in the data, that is patient or customer information that is not going to change during the analysis period, such as age, clinical or commercial records etc.

At the same time a new framework of analysis referenced as Recency-Frequency-Monetary (RFM), presented in more detail in section A.2, instead introduced a different scope to the study of survival analysis in the guise of churn prediction (section A.4) [54][23]. [15] defined RFM as:

- R (Recency): the period since the last purchase; a lower value corresponds to a higher probability of the customers making a repeated purchase.
- F (Frequency): number of purchases made within a certain period; higher frequency indicates greater loyalty.
- M (Monetary): the money spent during a certain period; a higher value indicates that the company should focus more on that customer.

The prototype of a very loyal customer thus presents low recency, high frequency and high monetary value. By employing metrics such as recency, frequency and monetary, the dataset is made up of observations of purchase metrics taken at periodic intervals. In this way the model can appreciate the evolution of customers' behaviour over time and draw not only from immutable, static features, but also from the dynamic evolution of the phenomenon, in this case the willingness of customers to place a purchase.

Most recent works and those that can be regarded as State of the Art in the survival analysis field [29][66][52][99] have finally started approaching it through sequential Deep Learning (DL) architectures, such as RNN, able to make predictions based on series of observations instead of static datasets. Such solutions often use attributes that can be regarded as *dynamic*, which means that their change over the analysis window allows the model to better understand the evolution of the phenomenon. Examples of such attributes can be the different observations of temperature gathered on various components of jet engines like in [66], or different diagnoses and analyses performed over time in a medical application, users recurrent musical interests like in [52] and finally features related to location check-ins like in [99]. Static features are basically ignored in such studies, even though error scores are not affected.

Table 1.1 shows a synthesis of all mentioned papers dealing with the task of surviving analysis, the techniques and models employed. More detailed descriptions on some of the ML techniques reported can be found in Appendix A.1.

1.2. Problem

As shown in the the Background Section 1.1, the investigation related to survival analysis through ML has mainly revolved around two main pillars: the nature of the utilised data (either static or dynamic) and the kind of architecture utilised (either sequential or non sequential). The possible combinations of these two characteristics have not been fully investigated, with the existing literature only focusing on two configurations. It can be observed that studies like [54], [90], [6], [45], [29] and [50] made use of exclusively static features while employing non sequential models. To the complete opposite instead there are approaches like [66], [52] and [99] that employed exclusively dynamic features on a sequential model.
This has been due to the more recent inception of sequential model compared to non sequential ones and the main consequence is that studies of possible combinations like static features used on sequential models or dynamic and static features used together on a sequential model are, to the best of our knowledge, almost non existent.

As shown by [28] the combination of static and dynamic features has in fact proven to be an interesting investigation in the context of RNN, showing how recurrent networks can outperform models like Logistic Regression and TLE [27] with a dataset composed of both dynamic and static features. Since the scope of this application is limited to two different medical tasks, the authors themselves stress the need for further experimentation. To the best of our knowledge, this kind of analysis has yet to be expanded upon regarding the churn retention domain, to which a case study could contribute in verifying the added value of combining static and dynamic features.

What emerges from the literature is that survival analysis applications can involve datasets with many datapoints for each user: [98][24][90][6][50] pertaining to applications with contract based services investigate datasets where each customer potentially has a datapoint for every day during the analysis window. This provides the model with a significant amount of events, that allows to employ rich feature representation of the investigated users. As shown in [58][101][12] there are also applications related to medical analyses and non-contract based business cases that offer less dense datasets. Each user presents only a handful of datapoints during the investigation period, which would cause overfit when the model is trained on many different features. Such an issue calls for the use of less complex, but efficient data representations. From the various formulations of data employed in survival analysis, the ones where RFM is involved [54][23][63] benefit from the fact that such a codification provides a simple representation of the data points with an intrinsic temporal nature. As it was introduced in [1] (and further referenced in A.1), the RFM framework can thus be a good choice when having to deal with applications where each user has a limited amount of datapoints.

Many survival applications, including all of the ones mentioned previously, limit themselves to provide a result as a binary classification: event happening or not. In churn prediction such a characteristic becomes quite a hindrance because, as stated by [96], there is a lack of direct operability on the results. Instead of just knowing whether a customer would churn or not, the authors applied clustering techniques to contextualize results to specific customer segments. This allowed to derive more detailed conclusions from the models. [36] showcases a case study also related to churn prediction in a fashion company. It also puts even more effort in making its results more actionable and better contextualized. This is achieved by conducting a preliminary clustering task on the dataset that identifies customer segments. As the author suggests, such a practice improved the performances of the different models, allowing to discuss results in a more fine grained manner and identify different behaviour on the single clusters.

This thesis focuses on the survival analysis problem in a fashion ecommerce setting. Com-

parisons on the error performances of different sequential and non sequential architectures will be conducted when trained and tested with two combinations of datasets: datasets composed of either dynamic only (RFM) features or a combination of static and dynamic features. Given the added value that a preliminary clustering task can bring to results interpretability, cluster information is included in the dataset.

1.3. Purpose

Given the introduction to the problem at hand, the research question that this study strives to answer to is:

Do static features improve the predictive power by reducing the error (Mean Squared Error (MSE)) in survival analysis problems?

1.4. Objectives

The study answers to the research question through the investigation of the survival analysis task on a transactional dataset made up of order information from fashion online retailer ASKET, spanning all orders between July 2015 and June 2022.

The study is going to contribute to the literature as a case study on the effects that the combination of static and dynamic features have on the performances measured in terms of MSE of sequential (LSTM) and non sequential (SVM, RF. xGboost and Multi-Layer Perceptron (MLP)) models. Furthermore it will try to make results more actionable and interpretable by performing a preliminary clustering task on the dataset.

These objectives can thus be summarised:

- Objective 1: Identify clusters in the available customer dataset, identifying meaningful and easily recognisable ones through K-Prototypes technique.
- Objective 2: Compare and analyse the results of the time to event predictions on sequential and non sequential models in terms of the contribution provided by static features in reducing the error

1.5. Methodology

This thesis proposes to answer to the reported research question by using an experimental research strategy [46]. By conducting experiments on several ML and DL models, most notably parallel tree boosters, MLP and LSTM, on a survival analysis task in two dif-

ferent data configurations, one including only dynamic temporal features and the other one including static features together with temporal ones (further referenced in 3.2.1), the difference in performances provided by the inclusion of said static attributes will be measured in terms of error. More specifically we are going to consider Mean Average Error (MAE) as the primary measure, as well as other indicators, such as MSE and R2-score. Previous applications of survival analysis such as this one have been already introduced in Section 1.1.
Before proceeding to the survival analysis itself a detailed analysis of the dataset will also be conducted in order to identify the most powerful and explaining features. By operating on it through K-Prototypes clustering technique (section 2.4.1) clusters will be identified from the customer base, that will be defined through features trends. This kind of information will then be included as a key static features in further analysis.

Using an abductive research approach [46], the results of the experiments will be analysed qualitatively to deduce an answer to the research question. In particular, the difference in performances on the various models between the dynamic only dataset and the one also including static features will give a clear indication of the advantages (and potentially disadvantages) provided by such an inclusion of data on the overall results of the prediction. The entity of such a difference will allow us to assess whether it is reasonable to consider such an inclusion in future tasks and further studies, for example given the improvement in the representation obtained by the combination of static and dynamic features. On the other hand, a marginal difference in performance will instead counsel against focusing on such combinations and instead open up towards further developments centred exclusively on temporal feature exploitation.

1.6. Possible limitations

- Although the dataset available for the study contains an extensive amount of customer orders, in many cases such customers have only interacted with the company once or twice. In order to make the analysis meaningful and have a reasonable amount of orders per customer to work on, only those with at least three recorded orders have been included in the analysis.
- The amount of static features to be included in the analysis is unfortunately not extensive as it could have been due to privacy policies not allowing to enrich datapoints with personal information about the customers. A more detailed insight about this will be given in the dataset overview 3.2.1.

1.7. Thesis outline

Chapter 2 (Theoretical background) provides an overview of the theoretical concepts required by the reader to fully appreciate the algorithms, models and metrics used in the thesis as well as a more direct insight in the task of churn prediction. Chapter 3 (Method) introduces the reasoning behind the choice of research method as well as the details of its application. It then proceeds with an introduction and overview of the data as well as some preliminary observations. Chapter 4 (Results) presents the results of the experiments. Chapter 5 (Discussion) proceeds to discuss on the numerical results, the gained insights and their interpretation. It also provides some indications of where the research can move on from the obtained results and what could prove useful to try. Chapter 6 (Conclusions) summarises the findings of the study.

Table 1.1: Synthesis of previous works.

Paper	Technique	Model
M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh [54]	Statistical approach on static features	Plain RFM framework
C.-F. Tsai and Y.-H. Lu [90]	DL	Artificial Neural Network (ANN) + Self-Organising Map (SOM)
C. G. Mena, A. De Caigny, K. Coussement, K. W. De Bock and S. Lessmann [23]	Comparison	Traditional ML vs RNN
N. Alboukaey, A. Joukhadar, and N. Ghneim [6]	Comparison	Statistical vs DL
N. Jain, A. Tomar, and P. K. Jana [50]	Multi Attribute Decision Making (MADM)	ML
C.-L. Hwang and K. Yoon [45]	ML	Logit Leaf Model (LLM)
P. Fader, B. Hardie, Y. Liu, J. Davin, and T. Steenburgh [29]	Statistical approach on static features	Beta-geometric distribution
E. Martinsson [66]	Statistical approach on dynamic features	Weibull Time To Event Recurrent Neural Network (WTTE-RNN)
H. Jing and A. Smola [52]	DL on dynamic features	LSTM
C. Yang, Y. Cai and C. Reddy [99]	DL on dynamic features	RNN
Y. Xie, X. Li, E.W.T. Ngai, W. Yingc [98]	ML on static features	Random Forest (RF)
E. Domingos, B. Ojeme and O. Daramola [24]	DL on static features	Deep Neural Network (DNN)
I. Kononenko [58]	ML on dynamic features	Naive Bayes Classifier
B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck and I. Bratko [101]	ML on dynamic features	Naive Bayes classifier
R. Bellazzi, B. Zupan [12]	Comparison	Decision Tree (DT), Logistic Regression, ANN, Support Vector Machine (SVM), Naive Bayes classifier, K Nearest Neighbours (KNN)

2 | Theoretical Background

2.1. Survival Analysis

We will now briefly digress on the topic of survival analysis to provide the reader with the required tools necessary to understand the forthcoming considerations and be able to follow the discussion. Most of this introduction draws from the detailed survey on this topic by [95].
Survival analysis is a subfield of statistics where the goal is to analyse and model the data where the outcome is the time until the occurrence of an event of interest. One of the main challenges in this context is the presence of instances whose event outcomes become unobservable after a certain time point or when some instances do not experience any event during the monitoring period. Such a phenomenon is called censoring and it makes the application of predictive algorithms through either standard statistics or ML techniques problematic [56]. Censoring can be divided in three main categories:

- right-censoring, for which the observed survival time is less than or equal to the true survival time;
- left-censoring, for which the observed survival time is greater than or equal to the true survival time;
- interval censoring, for which we only know that the event occurs during a given time interval.

It should be noted that the true event occurrence time is unknown in all the three cases. In most real-world cases we can also observe a prominence of right-censoring compared to the other two categories, like for example in the aforementioned applications in treatment failure assessments or business cases like evaluation of risks in insurance companies.
In survival problems we have absolute knowledge about the true time to event of interest (T) only for instances where the event occurred during the study period. For other instances we can only observe censored time (C) given that an observation may be lost. Those instances are considered as censored. That means that for any given instance i, we

can only observe one out of survival (Ti) or censored time (Ci).

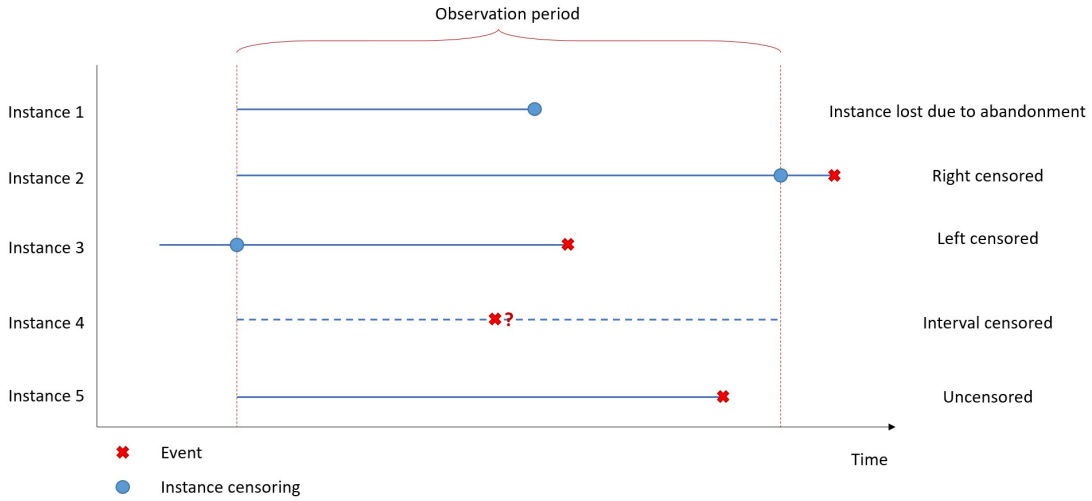


Figure 2.1: Examples of censored data.

In 2.1 we can observe a representation of censored instances: Instance 1 is a censored one due to the lack of information after the data point abandons the study or the relative information gets lost. Instance 2 is instead right censored since we lose the information after the end of the study, including the occurrence of the event. Instance 3 is left censored since we lose a portion of the observations happening before the studied interval. Instance 4 is instead interval censored since we only know that the event occurred somewhere in the studied interval. Finally instance 5 is an example of an uncensored data point, where we have full knowledge of all possible observations, including the final event.
From this point on we will refer to censored datapoints as right-censored ones, given they are the most common in most use-cases.
Let's now move to the specific formulation of a survival analysis problem. The survival function, also referred to as *S(t)*, gives the probability that a subject survives beyond time t. T is a continuous random variable with *F(t)* as its Cumulative Distribution Function (CDF). According to 2.1, the survival function is constantly decreasing in the range from 1 to 0 as t goes from 0 to 1.

$$S(t) = Pr(T \geq t) = 1 - F(t) = \int_t^{\infty} f(x)dx \tag{2.1}$$
$$0 \leq t < \infty$$

The hazard function λ in 2.2 denotes the failure rate at which the studied observations experience the event. The rate may increase or decrease as t increases. A hazard function

must satisfy the two conditions of 2.3 and 2.4.

$$\lambda(t) = \lim_{dt \to 0} \frac{pr(t \leq T < t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{2.2}$$

$$\forall u \geq 0, \lambda(u) \geq 0 \tag{2.3}$$

$$\int_0^\infty \lambda(u)du = \infty \tag{2.4}$$

The cumulative hazard function denoted Λ describes the failure distribution and can be expressed through 2.5. It is a non-decreasing function.

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.5}$$

The survival function and hazard function are closely related and a convenient formula can be derived from parts of 2.1 and 2.5.

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} log S(t) \tag{2.6}$$

$$S(t) = exp(-\int_0^t \lambda(u)du) = e^{-\Lambda(t)} \tag{2.7}$$

Some of the most popular survival analysis statistical methods are Kaplan-Meier approach [57], Log-Rank [87] and Cox Proportional Hazard Method [100], described in more detail in section A.3, but external to the scope of the thesis.

2.2. Machine Learning concepts

ML as defined by Mitchell is a branch of Artificial Intelligence (AI) where: a computer program is said to learn from Experience E with respect to some class of tasks T and performance measure P, improving with experience E.
The inference process performed by ML is induction as it tries to infer general rules from data and paradigms that are progressively discovered. The main difference of this approach compared to traditional programming is exemplified by the fact that in the first case, given the input and the code composing the program, the computer will provide an output for the task, whereas in the latter case the computer is provided with input and output, producing the program performing the task as a result. In essence ML aims at

replacing the development of software simply by providing the data.
ML by now has application in many different fields, comprising computer vision, speech recognition, biology, medicine and finance.
ML has many different subfields: supervised, unsupervised and reinforcement learning:

- Supervised Learning aims at learning the model, that is the correlation between input and output data
- Unsupervised learning aims at obtaining a better representation of the data since we have no indication of the desired output. A common task is dimensionality reduction, already introduced in section 2.3.
- Reinforcement Learning aims at learning how to control, to take actions based on experience.

Supervised Learning estimates the unknown model that maps known inputs to known outputs, with the notion of training set $D = \{\langle x, t\rangle\} \Rightarrow t = f(x)$.
Possible applications are varied and diverse, depending on the nature of the target t:

- if t is discrete, classification
- if t is continuous, regression
- if t is a probability of x: probability estimation

By approximating a function f given the input data D, a supervised model requires three elements:

- Loss function L: a function measuring the distance between any function and the to approximate. It tells how good or bad is the found function.
- Hypothesis space H: useful to restrict the attention to a subset of possible functions. It represents the insertion of human knowledge in the problem. Preferably it should be tailored to be small and very near to the true function.
- Optimization component h: the function to be found will minimise the loss function.

F is the space where the possible function f can be located. Furthermore hypothesis spaces can be expanded, but this usually increases the noise in the empyrical loss function because more function can be evaluated while the number of samples stays the same.

As mentioned, solving a problem in ML involves the employment of a model, which can be very different. For example:

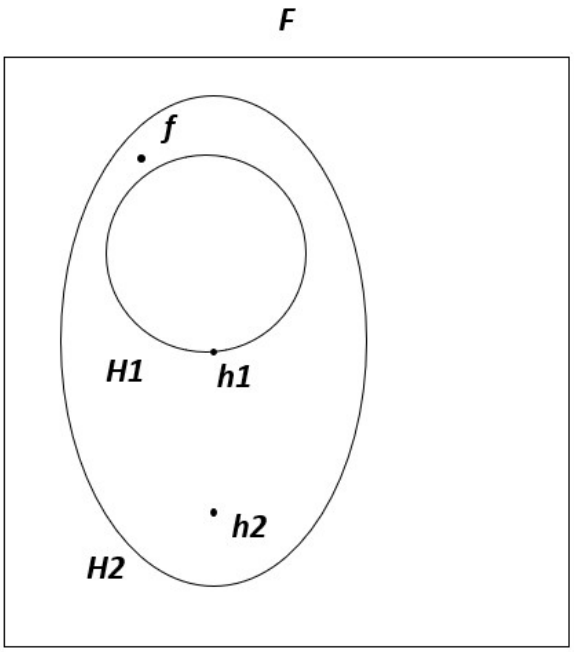


Figure 2.2: Supervised learning representation of the space of solutions.

- Decision trees: such trees are distinguished based on the nature of the target. Classification trees are used for discrete targets, while regression ones are used for continuous values. Decision trees can be represented as a series of decisions in a decision making process. All the input variables are divided in subsets, with some of them constituting the root of the tree and the other ones arranged in the internal nodes. External nodes (leaves) represent instead outputs and can either be of discrete or continuous form.
  It is also very common to build ensembles of decision trees:
  - Boosted trees: aimed at reducing bias it uses a series of simple decision trees sequentially b providing the full input to each of them and feeding the ouput of a tree to the following one. An example is AdaBoost.
  - Bagged trees: aimed at reducing variance, this technique build many different decision trees providing as input the same data with different replacements and uses voting to obtain a consensus from the trees. An example is Random Forest.
- SVMs are kernel-based methods characterized by being sparse, that is not utilizing the full extent of samples, but only subsets of them. Each sample is assigned a weight through which to include it in the analysis. SVMs represent one of the most complex algorithms in ML. They offer very strong guarantees and are exceptional

classifiers. They build on the loss function formulated for perceptrons: given a subset of samples X, a vector of weights a and a similarity function $K(x, x')$ kernel, a class prediction for a new example $x_q(t_i \in \{-1, 1\})$:

$$f(x_q) = sign\left(\sum_{m \in S} \alpha_m t_m k(x_q, x_m) + b\right) \tag{2.8}$$

where S is the set of indices of the support vectors.
Choice of kernel functions allow to project datapoints in highly dimensional representation spaces and they are chosen mainly leveraging on experience and knowledge of the problem. The main focus of the optimization is about working on the weights and examples. The optimization of weights follows the maximum margin criterion. This means that the separating margin between a point and the hyperplane has to be the one that is the most distant to the closest data points among the possible ones: $min_n t_n(w^T \phi(x_n) + b)$.
The maximum margin is found by solving:

$$w^* = \arg\, max_{w,b}\left(\frac{1}{||w||_2} min_n(t_n(w^T \phi(x_n) + b))\right) \tag{2.9}$$

The solution essentially maximizes the Euclidean distance of the point that is closest to the separating boundary (maximize a min problem).
Solving this problem is not trivial, thus it can be modified: what matters is defining the direction of the vector w, not its norm. Assuming a margin that is fixed at 1, the minimum can be changed to a new condition.

$$\text{Minimise } \frac{1}{2}||w||_2^2$$

$$\text{Subject to } t_n(w^T \phi(x_n) + b \geq 1, \text{ for all n}$$

The problem is solved by introducing a Lagrangian multiplier that transform the problem in an unconstrained optimization problem, where the constraints are encoded into the objective function:

$$L(w, \lambda) = f(w) + \sum_i \lambda_i h_i(w) \tag{2.10}$$

The optimal solution is found by obtaining: $\nabla L(w, \lambda) = 0$.
In this case we have to solve by considering conditions on both w and λ. In order

to solve it we need to apply KKT conditions:

$$\nabla L(w^*, \alpha^*, \lambda^*) = 0$$

$$h_i(w^*) = 0$$

$$g_i(w^*) = 0$$

$$\alpha_i \geq 0$$

$$\alpha_i^* g_i(w^*) = 0$$

The last condition is very interesting and produces two cases.
In the first case:

$$\text{Minimise } \frac{1}{2}||w||_2^2 = \frac{1}{2}(w_1^2 + w_2^2)$$
$$\text{Subject to } w_1 + w_2 \leq 1$$

The solution is in the global optimum, since the constraint is useless in this case.
In the second case:

$$\text{Minimise } \frac{1}{2}||w||_2^2$$
$$\text{Subject to } w_1 + w_2 \geq 1$$

The solution is no longer the global optimum, but lies on the constraint. We have that $g(w^*) = 0$ and $\alpha \geq 0$. When the solution is on the constraint, it is called active constraint. When a constraint is active, its Lagrangian multiplier is positive. On the other hand, if the solution is inside the region defined by the constraint, its Lagrange multiplier will be 0.
Every sample related to a constraint with positive Lagrange multiplier will be in the support vectors set.
Given the primal representation we have obtained, that is still parametric, we have to move to the dual representation by applying the condition that the gradient of L must be zero. We obtain:

$$\text{Maximize } \tilde{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m k(x_n, x_m)$$

$$\text{Subject to } \alpha_n \geq 0, \text{for } n = 1, ..., N$$

$$\sum_{n=1}^{N} \alpha_n t_n = 0$$

The classification of new points is done through:

$$y(x) = sign(\sum_{n=1}^{N} \alpha_n t_n k(x, x_n) + b) \tag{2.11}$$

When having to deal with noisy data, we may incur in the issue of a margin smaller than 1. Such cases are handled with the introduction of slack variables ξ_i. The primal formulation thus becomes:

$$\text{Minimise } \frac{1}{2}||w||_2^2 + C\sum_i \xi_i$$

$$\text{Subject to } t_i(w^T\phi(x_i) + b \geq 1, \forall i, \xi_i \geq 0, \forall i$$

What ξ_i does is shift the margin relative to sample x_i to respect the constraint. The term C manages how big the penalization is: with a high C, we are penalizing the misclassification a lot and they won't be allowed. On the other hand a C close to zero will barely penalise misclassifications, thus the resulting model will be very simple with high bias and variance.
The dual representation becomes:

$$\text{Maximize } \tilde{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m t_n t_m k(x_n, x_m)$$

$$\text{Subject to } 0 \leq \alpha_n \leq C, \text{for } n = 1, ..., N$$

$$\sum_{n=1}^{N} \alpha_n t_n = 0$$

C has become an upper bound for α_n and we can have three cases:

- $\alpha_n = 0$, the point associated to α_n is a support vector
- $0 < \alpha_n < C$, the point lies on the margin ($\xi_i = 0$)
- $\alpha_n = C$, the point lies inside the margin, and it can be either correctly classified ($0 < \xi_i \leq 1$) or misclassified ($\xi_i > 1$)

2.2.1. Deep Learning

In this day and age of technology and information, it does not surprise anymore to hear of the huge amount of data that becomes available every single second. While there are many and more applications constantly trying to make sense of such plenty, it still comes down

to the problem of how to exploit the data to the fullest, with as much of the process being automatised, in order to let it tell a story about the end results. [70] It is under such an influence that applications like ML, and a subfield of it, specifically DL, start prospering and benefitting by the ever increasing size of the data we can leverage. Thanks to this, ML can be regarded today as one of the most researched and prolific applications in the development of a variety of smart-world systems. We can find it employed in all kinds of industries and in relation to many different kinds of data, be they visual, audio, numerical, text, or some combination of them [55]. The extensive availability of such models and their capability to adapt to many different fields and applications really spells their success, especially when considered together with the fact that in many cases DL provides stark improvements when compared to previous applications [62][84][69].

The penetration of this new technique has also opened up new frontiers of development in some fields that began stagnating after initial improvements provided by ML. Just to name some of the most diverse applications, we can find automatic chord detection in what is referred to as music informatics, that poses itself the ambitious purpose of being able to efficiently detect chords from musical tracks in a matter of milliseconds [44][43], routines that make sure that we clean our teeth properly by using low power pre-trained models through data collected from sensor readings of electric toothbrushes, robots that flip perfectly cooked burgers, pour perfect beers or pollinate plants [4], autonomous systems that become progressively more precise and performing thanks to Reinforcement Learning (RL) [76]. Many different applications also exist in the health sector from medical image analysis [94] to system health management [55] and Alzheimer's disease diagnosis [8].

More formally DL is part of the broader ML family and focuses on representation learning and ANN.

2.3. Dimensionality reduction techniques

There are various techniques allowing to deal better with issues like the curse of dimensionality in data and the increased complexity and computation times it causes. Methods are often divided among linear and non-linear, while techniques can be classified in three categories:

- Feature selection: identifies a subset of input features that are most related to the output
- Regularization: all the input features are used, but the estimated coefficients are

shrunken towards zero, thus reducing the variance

- Dimension reduction: the input variables are projected into a lower-dimensional space

In best subset selections, the task is a combinatorial problem and we have 2^M possibilities, with M number of parameters. Given such a complexity a heuristic, as well as the risk of overfitting, has to be employed and those are the main options:

- Filter approaches: using some statistical tools the importance of features is found and ranked, selecting the best ones. Those methods are very efficient, but inaccurate since correlation between sample and features is not a relevant enough metric. It can either be a forward step-wise selection technique where the model starts empty and is populated by features a step at a time or backward step-wise selection where the model contains all features and at each step one (the worst one) is removed
- Embedded approaches: those algorithms solve the problem directly, while implicitly performing feature selection. Examples of such techniques are lasso, which puts weights of some features to zero, decision trees, auto-encoding etc.
- Wrapper: these techniques evaluate a subset of features.

Regularization techniques consist in the addition of components in the error function of problems that to various degrees tend to shrink the value of parameters to zero, thus reducing the overall amount. The most famous techniques are lasso and ridge regression that respectively introduce a l1-norm and l2-norm component as regularizer. SUch a reduction actively reduces the variance in the data.

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality [92]. Its aim to obtain a dimensionality reduced representation of the data that still preserves its intrisic dimensionality. So-called intrinsic dimensionality indicates the minimum set of parameters required to display the observable properties of the data [33]. Such a technique is fundamental in trying to address the issue of dimensionality reduction and the increased computational times that high dimensional representations entail [51].

The mathematical expression and meaning of dimensionality reduction can be thus expressed: let us consider a dataset organized in a n $\times$ D matrix $\mathbf{X}$ composed of n vectors x_i (i $\in \{1, 2, \ldots, n\}$) with dimensionality D. Considering the dataset having an intrinsic dimensionality d (d $<$ D, often d « D), that means that the points in X are placed on or in the vicinity of a manifold with dimensionality d that is embedded in the space. No assumptions are made on the structure of the manifold, which may be non Riemannian

due to discontinuities in the manifolds. The aim of dimensionality reduction techniques is to transform X (with dimensionality D) in a dataset Y (of dimensionality d), all the while preserving the data geometry. It is not known a priori neither what is the geometry of the manifold, nor the intrinsic dimensionality of X, d. It follows from this that dimensionality reduction is in itself an ill-posed problem, whose solution can be derived only by assuming given properties of the dataset (either d or the geometry) [92].

Traditionally dimensionality reduction problems have been addressed through solutions like Principal Component Analysis (PCA).

2.3.1. Principal Component Analysis (PCA)

PCA tries to construct a smaller representation of the data that retains as much information as possible. The idea is to find a linear basis, that is a set of 2 orthogonal axes, onto the projection of data retains as much of the orgiinal variance as possible. The projection of original data onto the first k principal components gives a reduced dimensionality representation of the data. This reconstruction will have some error, but it can be small.

In formulas, we can describe PCA as aiming at finding a linear mapping M maximizing the cost function trace, $M^T cov(X) M$, where cov(X) is the covariance matrix of data X. This linear mapping is formed by the k principal eigenvectors (principal components) of the sample covariance matrix. PCA solves the following problem:

$$cov(X)M = \lambda M \tag{2.12}$$

The eigen problem is solved by the d principal eigenvalues λ. The low-dimensional obtained representations y_i of the datapoints x_i are computed by mapping them onto the linear basis M, Y = XM.

PCA suffers from two main drawbacks. First, in PCA, the size of the covariance matrix is proportional to the dimensionality of the datapoints. As a result, the computation of the eigenvectors might be infeasible for very high-dimensional data. Second, PCA focuses mainly on retaining large pairwise distances, instead of focusing on retaining the small pairwise distances,which is much more important.

2.3.2. Factor Analysis of Mixed Data (FAMD)

Factory Analysis of Mixed Data (FAMD) is the factorial method devoted to data tables in which a group of individuals is described both by quantitative and qualitative variables.

In a context of dimensionality reduction, it works as a PCA for quantitative variables and as a Multiple Correspondence Analysis (MCA) for qualitative variables.

Such technique is required whenever the dataset under analysis is comprised both of numerical and categorical variables. In such an occurrence, the usual practice is to perform discretization of the quantitative variables. Data can thus be processed through MCA. Limitations to this are evident when:

- There are few individuals, which makes MCA unstable;
- There are too few qualitative variables compared to quantitative variables

Quantitative and qualitative variables are normalized during the analysis in order to balance the influence of each set of variables.

2.4. Clustering techniques

Cluster analysis is the task of grouping sets of objects in the same groups, called clusters, based on the property of them being more similar to each other than to elements present in other clusters. It falls down in the scope of exploratory data analysis and is in essence an unsupervised learning method.
For example, cluster analysis has been used to group related documents for browsing, to find genes and proteins that have similar functionality, and to provide a grouping of spatial locations prone to earthquakes. However, in other cases, cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbours of points. Whether for understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. Generally, clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods [73].

Partitioning or centroid-based clustering is characterised by having each cluster represented by a central vector, that is not necessarily a member of a data set. Whenever the number of clusters is set to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.
Hierarchical methods or connectivity-based clustering is based on the idea that objects are most similar to the ones that are closer to them. Clusters are formed by connecting objects that are close to each other. Clusters themselves can merge based on a given distance parameter and as such there is a hierarchy of clusters that can form and merge

based on the chosen distance parameter.
Density-based methods define clusters as areas with higher density of elements compared to the rest of the dataset. Elements that are located in sparse areas are instead considered as border points or outliers.
Grid-based methods are usually reserved for multi-dimensional data sets, by creating a grid structure constituted of cells. Such techniques are usually very light in computational terms and very fast.
Model-based methods is a statistical approach to clustering. The observed (multivariate) data is assumed to have been generated from a finite mixture of component models. Each component model is a probability distribution, typically a parametric multivariate distribution [9].

Cluster definitions themselves can differ to alternative categories:

- Exclusive clustering: each data can belong to only one cluster, so that there is no overlap among the various clusters. An example is the K-means algorithm
- Overlapping clustering: it uses fuzzy sets so that each data point may belong to two or more clusters, with different degrees of membership [68]. An example is fuzzy C-means clustering.
- Hierarchical clustering: such technique is based on the union of two of the nearest clusters into one [81], starting off with considering each data point as a single cluster. An example is the agglomerative algorithm.
- Probabilistic clustering: it uses a completely probabilistic approach. An example is a mixture of Gaussian algorithms.

2.4.1. K-Prototypes Clustering

While K-means is usually considered as one of the most common clustering technique, especially for large data, such a method is not suitable for use whenever the data set includes categorical variables. [41] proposed an algorithm called K-Prototypes able to deal with both numerical and categorical variables resulting from the notions of both K-Means and K-Mode.
What follows is the mathematical formulation as proposed in [42].
Supposing that $X = \{X_1, X_2, ..., X_n\}$ is a set of n objects and $X_i = \{X_{i1}, X_{i2}, ..., X_{im}\}^T$ where m denotes the variables and i denotes the i-th cluster.

The general formula for the measure of similarity is as follows:

$$d(X_i, Z_l) = \sum_{j=1}^{m} \delta(x_{ij}, z_{lj}) \tag{2.13}$$

Where $Z_l = \{z_{l1}, z_{l2}, ..., z_{lm}\}^T$ is a prototype for cluster l. A measure of similarity for numerical variables is well-known as euclidean distance that is denoted as follows.

$$d(X_i, Z_l) = \sqrt{\sum_{j=1}^{m_r} (x_{ij}^r - z_{lj}^r)^2} \tag{2.14}$$

Where x_{ij}^r is a value of numerical variables j, z_{lj}^r is the average of prototype for numerical variables J cluster m, and number of numerical variables.
A measure of similarity for categorical variables is instead denoted as follows:

$$d(X_i, Z_l) = \gamma_l \sum_{j=1+i}^{m_c} \delta(x_{ij}^c, z_{lj}^c) \tag{2.15}$$

While simple matching similarity measure for categorical variables is denoted as follows.

$$\delta(x_{ij}^c, z_{lj}^c) = \begin{cases} 0, & x_{ij}^c = z_{lj}^c \\ 1, & x_{ij}^c \neq z_{lj}^c \end{cases} \tag{2.16}$$

Where γ_l denotes the weight for categorical variables for cluster l that is standard deviation of numerical variables in each cluster. The X_{ij}^c indicates the categorical variables, z_{lj}^c is the mode for variables j in cluster l, and m_c is the number of categorical variables.
The modification of simple matching similarity measure is as follows.

$$\delta(x_{ij}^c, z_{lj}^c) = \begin{cases} 1 - \omega(x_{ij}^c, l), & x_{ij}^c = z_{lj}^c \\ 1, & x_{ij}^c \neq z_{lj}^c \end{cases} \tag{2.17}$$

The above formula increases similarity within clusters with categorical variables so that the result will be better where $\omega(x_{ij}^c, l)$ denotes the weight for x_{ij}^c where

$$\omega(x_{ij}^c, l) = \frac{f(x_{ij}^c | c_l)}{|c_l| \cdot f(x_{ij}^c | D)} \tag{2.18}$$

Where $f(x_{ij}^c | c_l)$ is the frequency of x_{ij}^c in cluster l and $|c_l|$ is the number of objects in cluster l, and $f(x_{ij}^c | D)$ is the frequency of x_{ij}^c in the whole data set.

According to 2.13 and 2.17, the measure of similarity prior to the data with numerical and categorical variables is as follows.

$$d(X_i, Z_l) = \sqrt{\sum_{j=1}^{m_r}(x_{ij}^r - z_{lj}^r)^2 + \gamma_l \sum_{j=1+i}^{m_c} \delta(x_{ij}^c, z_{lj}^c)} \tag{2.19}$$

The cost function declared by Huang for mixed types data sets is as follows.

$$Cost_l = \sum_{l=1}^{k} u_{il} \sum_{j=1}^{m_r}(x_{ij}^r - z_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} u_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, z_{lj}^c) \tag{2.20}$$

$$Cost_l = Cost_l^r + Cost_l^r$$

Where $Cost_l^r$ indicates the total cost of all numerical variables for the objects within cluster l. $Cost_l^r$ is minimized while z_{ij} is calculated as follows.

$$z_{lj} = \frac{1}{n_l} \sum_{i=1}^{n} u_{il} \cdot x_{ij} \text{ for } j = 1, 2, ..., m \tag{2.21}$$

Where $n_l = \sum_{i=1}^{n} u_{il} \cdot x_{ij}$ is the number of objects within cluster l.
Furthermore, the categorical variables like C_j area set of unique values in each categorical variable j and $p(q_{ij}^c \in C_J|l)$ is the probability for c_j within cluster l. $Cost_l^c$ can be thus rewritten as:

$$Cost_l^c = \gamma_l \sum_{j=1}^{m^c} n_l(1 - p(q_{ij}^c \in C_J|l)) \tag{2.22}$$

In order to minimise $Cost_l^c$ we can use the following.
For a specific cluster l, $Cost_l^c$ can be minimized if and only if $p(z_{ij}^c \in C_j|l) \geq p(c_j \in C_j|l)$ for $z_{ij}^c \neq c_j$ to all categorical variables. So that the cost function can be rewritten as follows.

$$Cost = \sum_{l=1}^{k}(Cost_l^r + Cost_l^c) \tag{2.23}$$

$$Cost = \sum_{l=1}^{k} Cost_l^r + \sum_{l=1}^{k} Cost_l^c \tag{2.24}$$

$$Cost = Cost^r + Cost^c \tag{2.25}$$

Since $Cost^r$ and $Cost^c$ are non-negative, $Cost$ minimalization can be done by minimizing the $Cost^r$ and $Cost^c$.

Through the mathematical formulation we are then able to use a method that performs just as well as K-Means on mixed data sets.

2.4.2. Elbow Method

As mentioned for K-Means and K-Prototypes in section 2.4.1, the optimal number of clusters has to be pre-determined in order to apply the algorithm as an optimization problem. In order to answer to such a need, the so-called elbow method is used.

The elbow method is a heuristic dedicated to identifying the optimal number of clusters in a data set. It consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.
The explained variation can be of varying nature and most commonly it is identifed with the variance. The aim is to choose a number of clusters so that increasing them would not provide a substantial increase in explained variation of the data set. In clustering terms, this means that the identification of an additional cluster would not provide much better modeling of the data.

3 | Method

3.1. Choice of Research Method

While trying to answer the research question reported in 1.3 there are various possible research approaches that could be employed, thus we will now present the most relevant ones along their strengths and weaknesses and finally the reasons as to why one has been preferred over the others.

One of the first choices to make when approaching the research question is choosing whether to use a theoretical approach or an empirical one. A theoretical research method mainly focuses on analysing previous research with the aim to derive considerations and further development in the field of study. In the case of this particular study it could focus on the results obtained by previous research with respect to the task of survival analysis and try and compare the results obtained in terms of error performances provided by the utilisation or lack thereof of static attributes in the datasets. The main advantages of such an approach reside in the fact that conclusions are gathered from various different datasets and applications, that is less influenced by contextual factors and more general in the perceived validity. On the other hand empirical research is heavily grounded on first hand experience through new experiments and data collected with the purpose of trying to validate directly any of the formulated hypotheses. In the context of the current research a theoretical approach to the problem is not preferable due to the fact that a paper directly addressing the comparison of performances with and without static attributes in survival analysis prediction is lacking, to the best of our knowledge. Most of past studies mainly focused on either utilising static attributes, or showing how using time varying features applied with models able to handle sequential data actually performed better than other standard solutions. Because of such inconsistencies, answering the research question by utilising a theoretical approach would have inevitably resulted in hardly defensible results. By employing an empirical approach instead makes it possible to compare the performances of the models in specifically developed testing conditions and obtain more reliable finding pertaining to the research question. Based on such reasoning, this research is going to be based on an empirical research method following

an experimental research approach. Such an approach is referred to as defined in [46], where it is defined as the required approach whenever one aims at establishing cause-effect relationships between observations and results.

As per the definition of an empirical research the answer to the research question can either come in the form of a quantitative or qualitative approach. Where a qualitative study would generally focus on the advantages and disadvantages brought by the inclusion of static attributes in the dataset, a quantitative one would instead be able to provide more tangible results via the set up of a benchmark on which to evaluate numerical results. Qualitative studies generally benefit by a wider array of attributes that can be included in the analysis, whereas a quantitative one would instead support each drawn conclusion with verifiable and quantitative results, making them harder to confute. In this study we will focus on utilising a quantitative approach given the factuality of the research question, that wants to provide a direct measure of the error as proof of a better reliability of the model. Along the just mentioned advantage, it also follows that such numerical comparisons are easier to interpret and again, harder to discuss against.

Following the framework provided by [46] to introduce a research methodology, let's briefly mention that in studies employing empirical research strategies, the best way of collecting data on which to draw conclusions from is experiments. It follows that such experiments should be set up by training models able of performing survival analysis predictive tasks and collect quantitative results on their performances and errors when using either purely dynamic temporal features or datasets also including static ones.

As in other regression problems, we can reasonably assess the performance of the models by measuring the MAE of its predictions based on transactional data. Given the intrinsic temporal nature of the data, its use demands for test sets to be drawn from observation gathered in a period following that of data used for model training. It is still possible though to divide the available data in distinct folds in order to perform cross validation and obtain a better indication of the model's predictions error.
Finally the data collected from experiments could be analysed either quantitatively or qualitatively. The former requires to employ a rigorous statistical methodology that can very clearly provide us with a definitive answer to the research question, assuming that we satisfy the requirements on the amount of data. The latter instead would configure itself more as a descriptive statistic, where we analyse the results and discuss about their relevance, while avoiding to investigate the precise causalities. Such an analysis, although being possibly opinionated and more prone to being confuted, does not require the extensive amount of data and experimenting of a quantitative approach. Given the difficulty of gathering enough data this study will mainly focus on a qualitative data

analysis to answer to the research question.

3.2. Application of Research Method

The following chapter will mainly address a thorough description and preliminary analysis of the available data, as well as an indication of the employed dimensionality reduction techniques and the feature obtained by the clusterization process conducted on the customer base, that provides better domain knowledge.

3.2.1. Data description

The dataset for the survival analysis task is provided by ASKET and consists of transactional data belonging to customer orders placed on the company's website between July 2015 and up until June 2022. It is composed of attributes belonging to varied and diverse aspects of the customer's experience and has been purposely fitted not to include any sensitive information about the customers. To exclude irrelevant or incomplete features cleaning the data takes a great part of the pre-processing stage to obtain a proper dataset for analysis. The customers in this dataset are limited to being users of ASKET and thus having created an account on the company's website. This results in a dataset consisting of 143922 customers who have placed 250324 purchase orders. Out of all the available features the need to identify the most relevant ones has resulted in the most informative ones as well as new ones obtained with feature engineering: *Exchanges Per Order* and *Average Order Window* are examples of such engineered features. Such a process will be showcased now.
From the initial assessment of the available data, a total of 35 features (2 of which are identifiers for order data and web interaction data respectively, *Customer ID* and *Segment ID*) have been extracted. Those features mainly pertained to 3 topic areas and are listed as such:

- Orders and returns data: features strictly regarding order's value, shipping, destination, purchased item categories and returned items
- Order interaction data: features loosely linked to orders including for example support tickets, review data and return reasons
- Web interaction data: all features extracted from website usage and related to customer's behaviour on the website, from the number of initiated checkouts, to the number of sessions conducted on average.

All features in Table 3.1 are showcased as aggregates computed over all the available orders

of a customer. This choice has been made in order to facilitate the initial clustering task as well as preparing the dataset for use in most of the ML models unable to work with sequential data. The clusterisation task, as referred to as in section 1.4, will provide us with more easily recognisable clusters through which customers will be referenced as well as providing an additional key feature in the analysis. The version of the dataset arranged as sequential will instead contain the same features as the static one without the aggregation (averaging) over the analysis period.

As a final clarification before delving deeper into the data presentation, it is worth mentioning that all data attributes and features pertain to a fashion retail store. As such there will be many references to garment kind, sizes and shipping methods. While such aspects might be unknown at the beginning, they are all thoroughly explained in this section and should be easily understandable even to a non expert audience.

Table 3.1: Initial data configuration.

Orders and Returns	Order Interactions	Web Interactions
Market	Most Common Return Type	Total Sessions
Most Common Payment Method	Most Common Category 1	Pageviews Per Session
Share Of Other Products	Most Common Category 2	Sessions Containing Cart view
Days from Last Purchase	Most Common Category 3	Number of Products Clicked
Number of Orders	First Purchase Source	Sessions Containing Purchase
Number of Items Purchased	Number of Tickets	Most Common Order Device
Number of Items Returned	Sizefinder Used	Sessions In the Last 6 Months
Item Level Return Rate	Average Review Score	Average cart Views
Net Sales	Coupon Used	
Most Common Delivery Method	Average Sentiment Score	
Paired Purchase Order Share	Predicted Gender	
Average Order Value	Most Common Collection	
Exchanges per Order		

A brief description of the meaning of each feature will now follow in Table 3.2:

Variable Name	Description	Type
Market	The country or geographical region where orders are placed from	Categorical
Most Common Payment Method	The payment solution chosen for the order	Categorical

Share of Other Products	The ratio between the number of orders containing an item belonging to the Other category and the overall number of purchase orders	Numerical
Days from Last Purchase	Number of days since the last purchase was placed	Numerical
Number of Orders	Overall number of purchase orders placed in the analysed window	Numerical
Number of Items Purchased	Overall number of items purchased in the analysed window	Numerical
Number of Items Returned	Overall number of items returned in the analysed window	Numerical
Item Level Return Rate	Ratio between number of returned items over the number of purchase orders in the analysed window	Numerical
Net Sales	Defined as the value of all purchased items minus the value of returned items in the analysed window	Numerical
Most Common Delivery Method	Chosen delivery solution	Categorical
Paired Purchase Order Share	Ratio of order containing items of the same kind but different sizes over the number of purchased items in the analysed window	Numerical
Average Order Value	Average of net value of all orders placed in the analysed window	Numerical
Exchanges per Order	Ratio of exchanged items over purchase orders in the analysed window	Numerical
Most Common Return Type	Reported solution chosen for a returned item	Categorical
Most Common Category 1, 2, 3	Most Common item category of the order as per in store classification, ordered by relevance (if at all applicable, like when there are less than 3 different categories purchased)	Categorical
Sizefinder Used	Flag on whether the customer has ever used Sizefinder, a built-in tool to provide accurate predictions on garment size with body measurements	Boolean

Average Review Scores	Average of review scores by the customer on purchase orders placed in the analysed window	Numerical
Coupon Used	Flag on whether the customer has used the one-time coupon on free shipping	Boolean
Average Sentiment Score	Average of sentiment scores associated to reviews by the customer on purchase orders placed in the analysed window	Numerical
Number of Tickets	Number of support tickets submitted by the customer in the analysed window	Numerical
Total Sessions	Number of overall navigation sessions by the customer in the analysed session	Numerical
Pageviews per Session	Average number of pageview visits per session by the customer in the analysed window	Numerical
Sessions Containing Cart view	Number of sessions containing a "cart view" (visualization of the cart in the website) event	Numerical
Number of Products Clicked	Number of clicks on product pages in the analysed window	Numerical
Sessions Containing Purchase	Number of sessions containing a completed purchase in the analysed window	Numerical
Most Common Order Device	Device used to place a purchase	Categorical
Sessions in the Last 6 Months	Number of sessions in the last 6 months of the analysed window	Numerical
Average Cart Views	Average number of "Cart View" events in the analysed window	Numerical
Most Common Collection	Most common collection belonging of purchased item	Categorical
Predicted Gender	Prediction on the gender of the customer based on their e-mail	Categorical

Table 3.2: Explanation of dataset variables.

To reduce the number of distinct values in the categorical variables, some sub-categories have been identified:

- *Market*: as per the relevance of the markets in the overall amount of orders, ASKET considers 6 primary markets between Europe and America, while then aggregating

the remaining countries in either "Rest of Europe" (RoE) or "Rest of the World" (RoW).

- *Product Categories*: ASKET offers items belonging to many different categories. Those categories are maintained according to their relevance against the overall volume of orders. The following categories of product are aggregated in "Other": Outerwear, Swimwear and Knitwear.
- *Product Collections*: all items sold are either developed as men's or women's collection items, with items belonging to the Garment Care category classified as Unisex items.
- *Order Devices*: purchases are usually made from many different kinds of devices. As a rule of thumb they are distinguished between two main categories: Desktop devices or Mobile devices (smartphones, tablets etc.). Since the launch of the physical store in May 2021, orders can be also placed there, thus the inclusion of the category "Store".
- *Return Types*: return types are usually not indicated, thus the unknown category is necessary, otherwise they can be either refunded or exchanged for other items.
- *Payment Methods*: payment can be made in various solutions. They are all aggregated in either "Klarna" checkout or "Paypal", with orders retrieved and paid for in the physical store reported as "Store".
- *Delivery Methods*: since ASKET ships all over the world, it collaborates with many different delivery services. In order to make the feature easier to understand they have been aggregated in either "Express" or "Regular" delivery Methods, with "Other" when such a distinction is not available.

In Table 3.3 we can find the summary of the possible values of the listed categorical variables resulting from the previous description.

Following from the discussion about the benefits of including RFM variables found in section 1.2 and having chosen to include features inspired by such a framework in the analysis as to improve the temporal nature of data itself, we are now going to provide a further clarification of those variables along their inception and definition in Table 3.4.

The RFM variables will prove particularly useful in providing the sequential models like the LSTM with a deeper time-bound understanding of the data.

Table 3.3: Categorical variables possible values.

Market	Product Categories
Sweden	Trousers
Germany	T-Shirts
UK	Sweatshirts
US	Shirts
France	Underwear
Denmark	Other
Rest of Europe (RoE)	Accessories
Rest of World (RoW)	Garment Care

Product Collections	Order Devices	Return Types	Payment Methods	Delivery Methods
Men	Desktop	Exchange	Klarna	Express
Women	Mobile	Refund	Paypal	Regular
Unisex	Store	Unknown	Store	Other

Table 3.4: RFM variables reference and description.

Feature	Context	Description
Number of Orders	Frequency	The number of orders in a specific analysis window provides with the general indication of how frequent the customer is in placing orders
Days since previous purchase	Recency	The distance in time between orders provides a sense of the customer's general behaviour in placing purchases
Net Sales	Monetary	Net sales show us whether the customer tends to be a big or little spender

On a different note, the dataset adjusted for the sequential models used in the analysis will differ in some aspects from the one we have just showcased. While the number of attributes and their inherent meaning will not change, their formulation will slightly change in order to accommodate for the "evolving" nature of the data. In general terms all features that are computed considering data belonging to the entirety of the window, will instead be computed step by step, that is as aggregates comprising data available until the time of each timestamped event, the purchase orders. That means that the *Market* feature could potentially vary at different time steps such as in the example of Table 3.5:

As evidenced in Table 3.5 the categorical features will be automatically set to the highest

Table 3.5: Example of categorical attributes behaviour in sequential dataset.

Date of Order	Country of Purchase	Market value in dataset
2021-04-19	Italy	RoE
2021-05-13	Sweden	Sweden
2021-09-03	Sweden	Sweden
2021-09-19	Sweden	Sweden
2021-10-25	Italy	Sweden
2021-11-06	Italy	RoE
2022-01-31	Sweden	Sweden

occurring value and in case of ties, to the most recent one. The behaviour is essentially the same one also in case of numerical and boolean variables, which basically means that we will compute variables accounting for the data present in the window up until that specific order. The only exception to this general rule is in the feature "Days Since Last Purchase", which will instead only consider the immediately previous order to compute the number of days between its occurrence and the currently analysed order, disregarding other events or the limitation caused by the dimension of the analysis window.

Finally there are some special clarifications to be made regarding the availability of some of these features: although most orders' features have been collected from July 2015 onwards, there are some of them which instead have been collected starting later on in the company's life.
Here is a list of when those features have started to be collected:

- All features encompassed in the *Web Interactions* topic have been collected starting from 20-05-2021 and not all customers placing an order after that date have been associated with this data according to their cookie acceptance policies.
- SizeFinder is size prediction service available on ASKET's website, whose data has been collected from 09-11-2020 onwards
- Reviews and Sentiment scores from reviews have started to be collected from 09-11-2019 onwards
- Information about coupon usage starts to be available from 14-01-2021 onwards

What this means is that for the mentioned attributes we have useful data only for periods following the beginning of their collection and thus they may yield less relevance when trying to address the whole period of ASKET's activity in the prediction tasks.

3.2.2. Data mining

In order to achieve a less complex configuration of the available data. different dimensionality reduction techniques have been evaluated and tried on the data. Most importantly FAMD has been employed to try and obtain a reduced representation of the data. The obtained results have struggled to meet up with the required expectations and are thus not analysed, but further referenced in the Appendix section C for those interested.

Further analysis and comparisons in the correlation between similar features that had been identified in the dataset have followed, demonstrating some redundancies in the data. Such process is explained in detail in Appendix section C and has provided a final configuration of the data for the experiments, showcased in Table 3.6.

Table 3.6: Final data configuration.

Orders and Returns	Order Interactions	Web Interactions
Market	Most Common Return Type	Total Sessions
Exchanges per Order	Most Common Category 1	Pageviews Per Session
Days from Last Purchase	Most Common Category 2	Most Common order Device
Number of Orders	Sizefinder Used	
Number of Items Returned	Coupon Used	
Net Sales	Number of Tickets	
Most Common Delivery Method	Predicted Gender	
Average Order Value	Most Common Collection	

3.2.3. Survival analysis approach

After having got a thorough introduction and presentation of the data, its composition and its relevance to the purpose of the analysis, a small mention of how it will be best shaped for the task and any kind of limitations to the assessment is due.

In order to make the analysis worthwhile, especially in the context of ML models that cannot process sequential data, the initial dataset has been further reduced to include only customers that have purchased at least 3 orders in the analysed window on the example of [66]. This cleans the dataset from occasional buyers and feeds the model only with quality data showing actual customers' behaviour. When considering such a limitation, the dataset gets reduced to around 24k customers active between July 2015 and beginning of 2022.

3.2.4. Metrics

The analysis of the performances of models mainly relies on an inspection of the MAE based on two different data configurations. While this represents the main reference metric, there are others that are also going to help us get a grasp of the model's behaviour. They are MSE and R2-score. A brief presentation and explanation of the relevance to the experiment assessment will follow, as well as two added metrics that we will take into consideration:

- MAE: a measure of errors between paired observations expressing the same phenomenon calculated as the sum of absolute errors divided by the sample size [97].

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

- MSE: it measures the average of the squares of the errors and the quality of an estimator. As it is derived from the square of Euclidean distance, it is always a positive value that decreases as the error approaches zero [2].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2$$

- R2-Score: the coefficient of determination, denoted R2, is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model [39].

Given

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$SS_{res} = \sum_i e_i^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y})$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

The reasoning why MSE is preferred over MAE is that overall we want to make sure that also some of the outliers' behaviour gets optimized and taken into account, in order to have a behaviour consistent with even less common cases. Given that the MSE uses a squared error in its estimation it is only normal it would address these kinds of errors better. MAE in itself is instead a more generalised indicator that can still provide us with reasonable insights.
R2 Score instead is a useful indicator of the goodness of fit of a model, measuring how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data. This can be good to understand which data configuration or model provides a more "on-point" prediction.

Given the temporal nature of the data, there are two additional constructed metrics aiding in the interpretation of the model and in the approach to its error in the solution. They are respectively the average of the positive and negative difference between the predictions and the true value:

- Average positive offset: average calculated over all the predictions that are after the true value in terms of days

$$APO = \frac{\sum_{i=1}^{n}(TOE - TrueValue)_i}{n}$$

  if *TOE - TrueValue* > 0

- Average negative offset average calculated over all the predictions that are before the true value in terms of days

$$ANO = \frac{\sum_{i=1}^{n}(TrueValue - TOE)_i}{n}$$

  if *TOE - TrueValue* < 0

Such metrics help us in getting a more immediate sense of the general error committed by the model in its predictions and can prove useful in a business case when trying to define a window in which the probability of the event happening is very high, defined simply as [*TOE - Average negative offset; TOE + Average positive offset*].

3.2.5. Dynamic vs. dynamic + static data

One final aspect of the application of the research method pertains how the evaluation of model's errors will happen when considering only dynamic or temporal features and when such features are complemented by static ones as well. This is fundamental to the

ends of answering the research question and will develop by extracting two versions of the dataset from the one defined and analysed in sections 3.2.1 and 3.2.2.

Starting off from the list of features provided in Table 3.6, we will make a distinction between *dynamic* ones, that are going to make up the basis of the dataset and those that will instead be referred to as *static* ones, that will only be included in a single, extended dataset in addition to *dynamic* ones.
To keep things clear, *dynamic features* will be represented by RFM features, able to give a very good representation of the evolution of order placement behaviour from customers over time (as explained in section 1.1. *Static features* instead represent the majority of other information available to us, mostly order information aggregates and related metrics.

Table 3.7: Dynamic feature dataset composed only of RFM features.

Recency	Frequency	Monetary
Days until first purchase	Last purchase net sales	Number of orders

The configuration of data provided in table 3.7, while extremely simple and only complemented by identifiers, that are useless for the analysis and thus ignored, will effectively be able to give a vision of how each customer behaves on a *monthly* basis. Based on the initial limitations over data imposed by choosing customers only placing at least 3 orders in the analysis period, the experiment has been set up choosing as experiment period the richest and most apt one for the analysis. Training data will be comprised of all order information available for such customers on a 12-month period spanning from March 2021 and March 2022. The choice, while further limiting on the data, is dictated by the difficulty of creating a monthly dataset due to manual data retrieval and allowed for extensive experimentation over the various configurations.

An important note has to be made on the slight difference in how data is presented to the models based on their nature. As already mentioned in the Introduction chapter, the focus will be both on traditional ML models that are unable to work with timestamped data, like the dataset in table 3.7 is, and RNN models that instead have a hard requirement for such a configuration. The solution we adopted is to create, based on the same data, an "aggregated" version of the dataset that will essentially save all the RFM features of orders placed in the analysis period and present them together to the model instead of divided in a series of 12 snapshots.

In table 3.8 we can find the configuration of the dynamic features as provided to non sequential ML models.

Table 3.8: Dynamic feature dataset for models not able to deal with time varying models.

Recency	Frequency	Monetary
Days until first purchase	Last purchase net sales	Number of orders at last purchase
Days between last to one to last purchase	One to last purchase net sales	Number of orders at one to last purchase
Days between one to last and second to last purchase	Second to last net sales	Number of order at second to last purchase

A similar treatment is used for *static* features. In this case the features are aggregated over the twelve months for the non sequential ML models and instead provided on a monthly basis for sequential models.

Table 3.9: Static features to be included in a static + dynamic dataset.

Orders and Returns	Order Interactions	Web Interactions
Market	Most Common Return Type	Total Sessions
Number Of Items Returned	Most Common Category 1	Pageviews per Session
Average order grand total value	Most Common Category 2	Most Common Order device
	Sizefinder Used	
	Coupon used	
	Number Of tickets	
	Predicted gender	
	Most Common Collection	

In table 3.9 the list of static features is provided divided in the same fashion as they were presented in section 3.2.2.

4 | Results

In this chapter the results of the specific implementation of the research method showcased in chapter 3 will be presented, with indications of technical aspects and the end results. The section will provide an initial assessment of results related to the clusterization task described as the first step in the Objectives section, with a brief interpretation of the results and how this information will be included in the dataset for the survival analysis predictive task. A conjoined assessment of results will then follow in chapter 5.

4.1. Clustering

The first step in developing an answer to the research question has been to make sense of the multitude of data and the extreme variety of customers present in the dataset by employing a clusterization technique. This allows to abstract a more generalized and easily interpretable representation of different customer types in the dataset. It proves useful to the direct ends of assessing how a combination of static features may be useful to the performance of models in the survival analysis task since different clusters may represent represent behaviours in the data ranging between two extremes: on one side there may be customers presenting order behaviour (datapoints) that tends to stick to the general behaviour of other similar customer, while on the opposite side there may be the outliers and more erratic customers that instead may present evident differences in their order placement, but are still characterised by similar static features.
The added understanding of the dataset that such insights may provide is very beneficial to the ends of the final discussion and is now presented.

As mentioned before, the dataset is composed of mixed data, both categorical and numerical. Such a characteristic requires the use of a KNN-Prototypes technique able to deal with such diversity in the nature of data.
The first step the method is to decide how many clusters to identify. Such decision is supplemented by using the so called "Elbow Method", further referenced in section 2.4.1.

As can be observed in Figure 4.1 the behaviour of the clusterization cost curve is that of

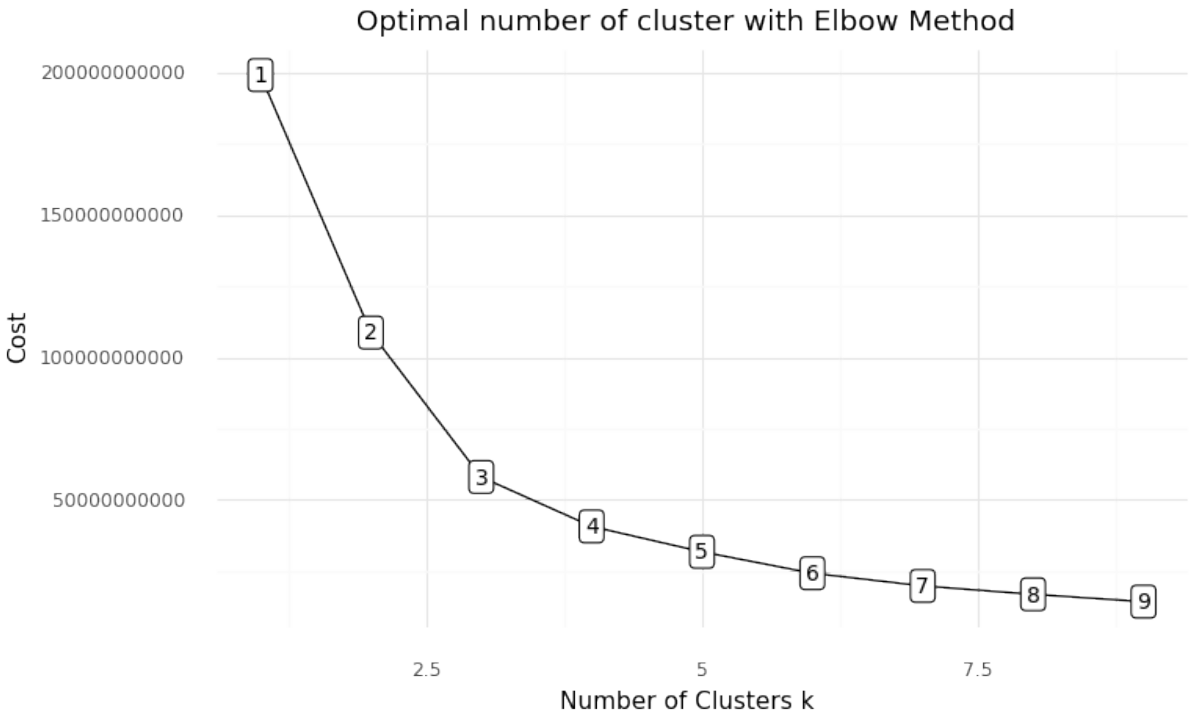


Figure 4.1: Cost of the clusterization task from 1 to 9 clusters using the Elbow Method.

a gradual reduction of added cost as the number of clusters increase. The way a number is chosen is based on the instance where the difference in cost of clusterizing the dataset with n or $n + 1$ clusters becomes very little. Through more detailed inspection of the clusters provided by employing $k = 4$ or $k = 5$, that is omitted for the sake of simplicity and only pertains an assessment of variance in the dataset itself, the choice falls on a number of clusters $k = 5$.

What follows now is a presentation of the main characteristics of the clusters and how they will be referred to from now on.

Only the most relevant features to our analysis have been included in table 4.1 and based on some domain knowledge it is possible to derive a more comprehensive interpretation of each cluster, as well as a way of defining them:

- Cluster 0: this cluster represents the oldest customers that have ever interacted with the company. As can be noted from the high value of days since last purchase they mostly represent some of the first people interacting with ASKET in a reduced capacity, that is by placing a very small amount of orders of little value and then shifting away from purchasing from the company altogether. This is further shown by the fact the main market is Sweden, the first one serviced by ASKET, as well as

Table 4.1: New cluster names.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Percentage	16%	8%	28%	48%	1%
Days since last purchase	Highest (at least 746 days)				Lowest
Average order value (SEK)	Low (1000)	High (4200)	Average (2200)	Low (1000)	High (9000)
Sizefinder usage	Low (not in the last 2 years)	Average	Average	High	Low
Most common category	T-shirt	Trousers	Trousers	T-shirt	Trousers
Market	Sweden	Germany	Germany	Sweden	Germany

not having interactions with many of the recent years services, like Size Finder and other web tracked activities on the website.
In terms of RFM they are the least frequent buyers, providing the smallest amount of revenue and rarely ever purchasing anymore. They feat neatly into the definition of *churned* by having completely terminated their interactions with the company.

- Cluster 1: this cluster is populated by customers showcasing a highly varied behaviour in terms of recency of purchase, although their frequency is high, usually sporting a number of orders in the higher end of the interval. They also show a decent amount of interaction with web tracked activities and Size finder, proving to be mostly recent or very recent customers. With the majority of them belonging to Germany, representing ASKET's second biggest market and one developed in recent years and months, they prove to be the backbone of the most loyal of ASKET's customers based on their higher profitability. This last aspect is reinforced by the tendency to buy Trousers, which represents a more expensive item than the T-shirts, and one introduced later.

- Cluster 2: this cluster represents a cohort of customers that share different aspects with Cluster 1, mainly in their recency and entity of purchase. The main differences reside first in the increased amount of time spent navigating on the website and a comparatively higher average of support tickets. This shows the increased indecisiveness of such a cohort, that still has to decide whether they trust ASKET's products in their purchases.

Going back to a RFM framework interpretation, they boast high recency, but lower frequency and monetary value than that, thus representing some very good "one-purchase customers" that still need to show whether they will stay loyal or not.

- Cluster 3: this cluster, representing the overwhelming majority of the whole customer base, presents quite a uniform behaviour contrary to expectation. Similarly to Cluster 0 it is mainly made up of older customers, given that the majority of them hail from Sweden and have mostly purchased T-shirts (the first item to ever be available on ASKET's website). This means that such customers already have a history with ASKET, have tried its products and could potentially come back from time to time for a new purchase, although their current purchase behaviour is less frequent than in past years. A reasonable argument to make is that they are the ones where the Time To Event (TTE) prediction will prove most useful, in order to make sense of their future intentions with the company. Overall they boast average recency, low monetary value but high frequency.

- Cluster 4: this final cluster, representing the smallest section of the customer base could of been easily excluded from the analysis, but it is useful to agglomerate most of the outliers in the dataset. It represents the small cohort of highest spenders that have started buying from ASKET at various times, but mostly recently, and are interested in the higher end products. They are a staple source of ASKET's income and probably some of the most consistent buyers, helpful to analyse the long term behaviour of ASKET's customer analysis.
  Overall they boast the highest in terms of frequency and monetary value, with recency being dependent of their own attitude to purchase, mainly because ASKET itself, as a slow fashion company, has a less frequent interaction lifecycle with their customers

Aside from the current discussion on cluster's characteristic, we can further appreciate a representation highlighting the clear distinction between clusters through their behaviours along the "Days until last purchase" and "Average order value" features, which represent evidently the most important ones the model took into account in the clusterization process.

As Figure 4.2 shows, most of the subdivision between clusters happens based on their usual purchase behaviour, and in particular the entity of purchases they usually place. As already mentioned in Table 4.1, ASKET'ers and Loyal Customers represent the most profitable components of the customer base, with the vast majority of them having quite small days until last purchase and higher than average spending.

On the other hand whenever the purchasing power is generally similar, the most important discriminating factor becomes the "Days until last purchase". This is how Churned and Potential Loyalists get differentiated.

Based on this extensive introduction and comparison between customers, it is possible to draw a final concluding view of the obtained clusters also by assigning them some more easily recognisable names. We can observe this in table 4.2 along the most relevant aspects.

Table 4.2: Cluster characteristics.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Churned	Loyal Customers	Indecisive Shoppers	Potential Loyalists	ASKET'ers
Old customers, not interested anymore	Frequent buyers, big spenders, already hooked with ASKET	Interested customers with low amount of orders. Could potentially become regular customers	Older customers who still interact with ASKET. Might be solicited to become more active again	Biggest, most frequent and recent buyers. Essentially brand ambassadors.

Now that the cluster analysis is sufficiently showcased it can be used to enrich the data representation by associating the cluster affilation to each order and customer in the TTE analysis. This will allow us to draw some more detailed and context-aware conclusions in the Discussion chapter and imbue some business knowledge to the benefit of the task. Specifically the new Cluster feature will be a Categorical static feature with 5 possible values.

4.2. Time To Event prediction

According to the second point in the Objectives section, this chapter addresses the results pertaining to the survival analysis task based on the data presented in chapter 3 and the additional information provided by the Clusterization task reported in section 4.1.

The task itself consists of assessing the capabilities of several ML and DL models in predicting TTE in a churn prediction task, which is the direct outcome of survival analysis according to its definition.
The models used belong to two different categories as they were introduced earlier:

- Non sequential ML models: such models are incapable of dealing with sequential data, requiring for sequential data aggregation in order to make the analysis feasible.

- Sequential models: those models have a built-in mechanism allowing them to retain knowledge from other sequences of data and as such are built in order to accommodate sequential data.

Table 4.3: List of models employed in the analysis.

Non sequential ML models	Sequential models
SVM	LSTM
RF	
xGboost	
MLP	

In table 4.3 a brief list of models employed in the analysis is provided.
Regarding the first category of non sequential ML models, the choice has fallen on many different models spanning from SVM, tree-based structures and a DL architecture that have been most commonly used in the literature (see section 1.1) to analyse the effect of static features on the error of a survival analysis.
Similar reasoning follows in the sequential section, where the LSTM is chosen and preferred over other RNN architectures because of the stark improvement it provides over them in terms of memory and predictive performance.

As commonly adopted for ML applications, models' performances and errors have been analysed through the use of cross-validation techniques.
In particular a training-test split of 70/30 has been applied in order to comfortably allow for a good portion of customers to be unseen to the model and thus provide as good an evaluation of the error (and eventual over or underfitting) as possible.
In regards to the validation set instead a further 20% of the dataset has been at each run dedicated to it, with a k=3 employed in the k-fold cross validation technique. This ensures a good optimization of the parameters.

Given the diverse nature of the models employed, a list of optimized parameters with possible values used in the training of the model will be listed with the final best performing ones properly highlighted. Besides that, empirical results will be presented altogether in a single table to allow for easy comparison between the models.

In table 4.3 we can observe the optimal parameters identified for the model in bold. In most cases the models do not need to be exceedingly large or complex, especially in DL models, given that the number of available samples is not very big, despite the variety of the data they represent.

In table 4.5 the error measurements and average offsets obtained from the various listed

Table 4.4: List of models and optimized hyperparameters.

Model name	Hyperparameter name	Possible values
SVM Regressor	C	[0.1, **0.5**, 1, 3]
	Epsilon	[0.5, **1**, 1.5]
Random Forest Regressor	N. of estimators	[50, 80, **100**, **150**]
	Max features	[auto, sqrt, 0.4, **0.6**, 0.8]
	Max depth	[None, 50, 100, **150**]
	Min. samples split	[2, 10, 30, 50, **70**, 90]
	Min. samples leaf	[1, 30, **50**, 70, 90]
xGboost	Eta	[1e-3, 1e-2, **0.1**, 0.3]
	Subsample	[0.3, 0.5, **0.7**, 1]
	Gamma	[0, 1, 5, 10, 20]
	Reg_alpha	[**0**, 0.2, 0.4, 0.6]
	Reg_lambda	[**1**, 1.5, 2, 3]
MLP	Learning rate	[1e-5, **1e-4**, 5e-4, 1e-3]
	Optimizer	[**Adam**, Adaboost]
	Early stopping epochs	[10, **20**, 30, ∞]
	N. of units	[128, **256**, 512, 1024]
	N. of layers	[1, **2**, 3, 4]
LSTM	Learning rate	[1e-5, **1e-4**, 5e-4, 1e-3]
	Optimizer	[**Adam**, Adaboost]
	Early stopping epochs	[10, **20**, 30, ∞]
	N. of units	[32, **64**, 128, 256]
	N. of layers	[1, **2**, 3, 4]

models when using a dataset composed only of dynamic features can be observed, as introduced in section 3.2.5. Through the use of just RFM variables it can be clearly observed that the LSTM model was able to provide a better all around predictive performance when compared to other models that are unable to handle data in a sequential way and had to rely on aggregations of the dynamic data.

The ability of the LSTM model to collect an evolving and time-bound representation of the data has definitely helped it in bulding a better abstraction of it and thus a better performance when handling predictions on unseen data. In particular we can observe a **14%** reduction in MSE, **8,4%** reduction in MAE and **20,3%** reduction in average negative offset when comparing the LSTM to the best performing non-DL model (all the traditional ML models except the MLP) for each metric.

It is also worthy of observation that the non-DL models were instead able to provide a less chaotic predictive performance according to the R2 score, probably because of the unnecessary complexity of the MLP and LSTM models compared to the task at hand. It is also because of this that such "less powerful" models were able to keep up in per-

Table 4.5: List of models errors and metrics with a pure dynamic dataset.

TTE prediction	SVM	RF	xGboost	MLP	LSTM
MSE	0,029	0,029	0,034	0,035	**0,025**
MAE	0,131	0,139	0,147	0,138	**0,12**
R2 Score	0,115	0,086	**-0,014**	-7,397	-1,55
Average offset	-79, +127	-94, +124	-102, +125	-63, +134	**-63, +124**

formances given the variety in the data could still be represented without unnecessarily complex abstractions.
This brief assessment can be complemented by mentioning that most of the errors committed by the various models are on the positive, that is contributing to the average positive offset. This seems to be a tendency of the DL models mostly and could be further investigated.

Table 4.6: List of models errors and metrics with dataset consisting of both dynamic and static features.

TTE prediction	SVM	RF	xGboost	MLP	LSTM
MSE	0,028	**0,026**	0,029	0,04	0,028
MAE	0,133	0,13	0,133	0,149	**0,121**
R2 Score	0.127	0.203	**0.117**	-8.54	-2.08
Average offset	-85, +123	-86, +121	-102, +125	-59, +144	**-49, +125**

In table 4.6 the error and performance measurements can be observed when evaluating the models with a combination of both dynamic and static features.
The error and metrics put an edge of traditional ML model over the LSTM, in particular: MSE sees a reduction of error of **7%** of the RF compared to LSTM, a reduction of **7%** in MAE and of **43%** in negative average offset by the LSTM compared to the RF.

Differently from before we can observe a different trend in performances: in terms of errors the traditional ML models seem to have benefitted most from the addition of static features in their performances. While the number of samples has not changed, the addition of static features has created a more complex and rich data representation that because of its nature has not benefitted in the same way the LSTM model, that does not deal well with unchanging data across different time steps.
The increased complexity of the data representation is evident when observing a shift in R2 score behaviour, although in this case the difference between values in the two configurations does not necessarily mean the model improved or not, but rather the absolute value the score assumes, which is generally higher in the second configuration, tells the

prediction tends to be more chaotic.
Even with all such difficulties, the LSTM model still manages to provide an overall best performance in terms of average offsets. This means that while the predictions may tend to be slightly less precise, when the model makes an error, the prediction usually tends towards the true value instead of going off in the opposite direction. There is argument to say the model itself has a better contextual representation of data compared to the non sequential models.

Table 4.7: Percentage changes between model performances when using only dynamic features and dynamic + static ones.

Variation	SVM	RF	xGboost	MLP	LSTM
MSE	-3,5%	-10,3%	-14,7%	+14,3%	+12%
MAE	+1,5%	-6,4%	-9,5%	+6,5%	+ >1%
Average offset	+7,6%, -3,1%	-8,5%, -2,5%	0%, 0%	-6,8%, +7,5%	-22,2%, + >1%

In table 4.7 it is possible to observe the difference between the errors and scores of the model in the configuration with static and dynamic data against the one with only dynamic data.
As already briefly mentioned, the inclusion of static attributes in the analysis seem to have provided different benefits to the various models and we can summarise such changes in three points:

- Non-DL models: For the SVM, RF and xGboost models the introduction of static features has aided the models in reducing their prediction errors, in particular in the cases of tree structures. This shows how a configuration including only dynamic attributes actually affects the predictive performances of such models negatively, since they are unable to process this kind of information effectively after the required aggregation and instead prefer working on features and attributes that are natively built to represent invariant, static information regarding the whole analysis window.
- DL, non-RNN models: in the case of MLP, while the model is not meant to handle sequential data, the introduction of static features did not improve the error performances, but rather made the model struggle with predictions even more. This is probably due to the increase of complexity in the prediction that may require better optimization on the number of samples compared to amount of features. DL models are in fact traditionally data hungry.
- RNN models: the LSTM models, similarly to the MLP, does not benefit from the introduction of static features, mostly because of the increased data complexity that

it causes and the lack of an increase of available samples at the same time. In general though, such reduction is less felt and it still manages to retain most of its predictive power and its best performance in terms of average offset, thanks to the reliance on the sequential configuration of data. Still the inclusion of static attributes can be regarded as the wrong step in an attempt to increase data complexity.

The specific implications of such results are very diverse and strictly dependent in each single category of models on the way such models themselves handle different kinds of data.

As a final note it might be interesting to observe the behaviour of predictions compared to ground truths in a cluster-wise environments. This will be shown regarding LSTM model results, as the best performing model out of all the considered ones.

As can be observed in figures 4.3 and 4.4 the prediction distribution between cluster, although different according to cluster-wise behaviour related to the inherent characteristics from customers of each one, still provide us with a general idea of how the model behaves and where its limitations lie.
As could be predictable, most of the predictions tend to be concentrated towards the average amount of days between purchases, with less and less predictions moving towards smaller or higher amounts. At the same time such a behaviour may be influenced by a general tendency of the model to accommodate for the general behaviour of customers and instead struggle with outliers within the same cluster.
When analysing the distributions regardless of cluster belonging though, it is possible to appreciate how the cluster information clearly helps the model in identifying which cohort of customers each analysed data point belongs to and from that provide a more reliable prediction based on this additional information. It is safe to say that without encoding such static information the performances would of definitely been worse, but this is outside the scope of this investigation.

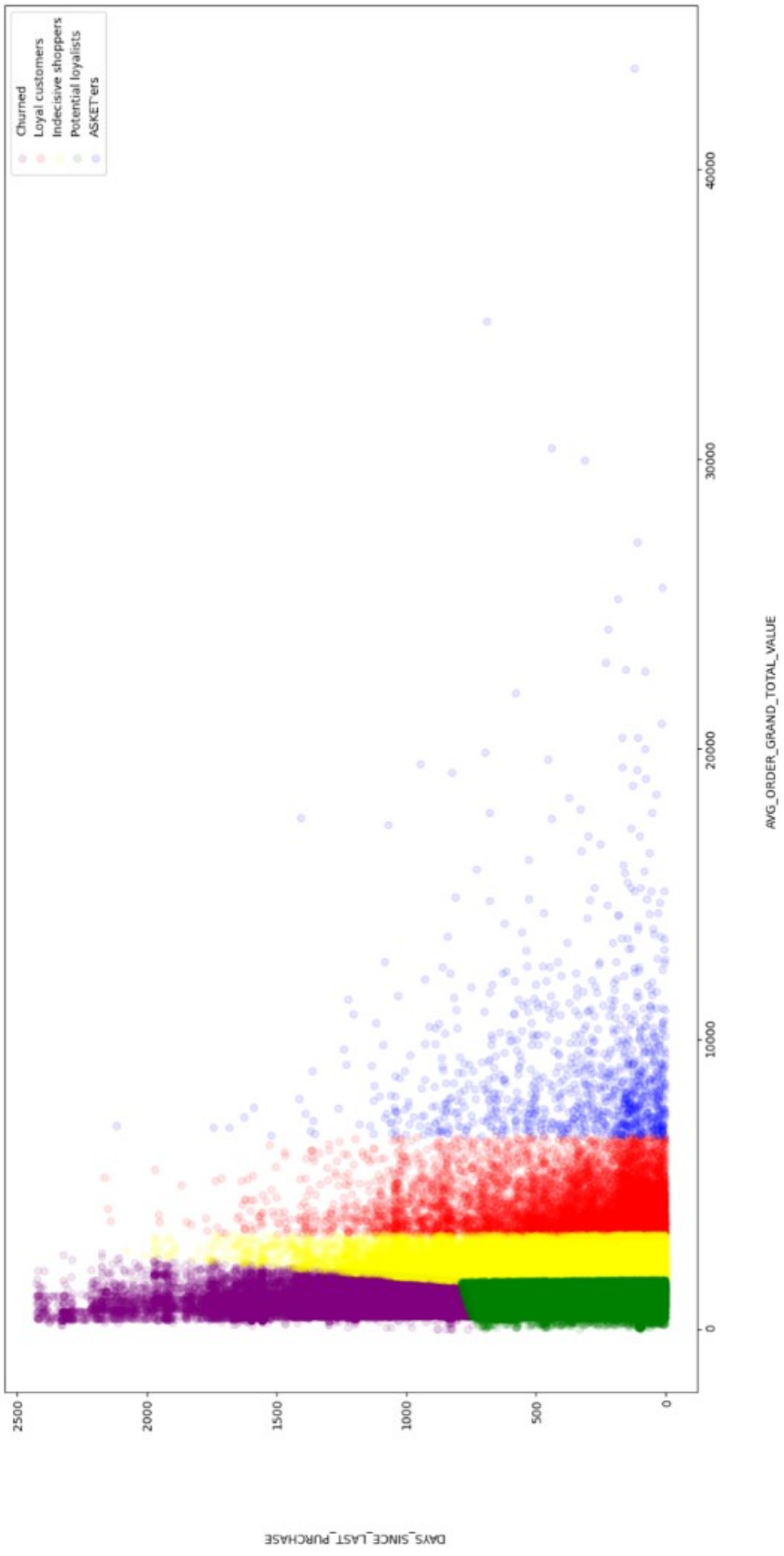
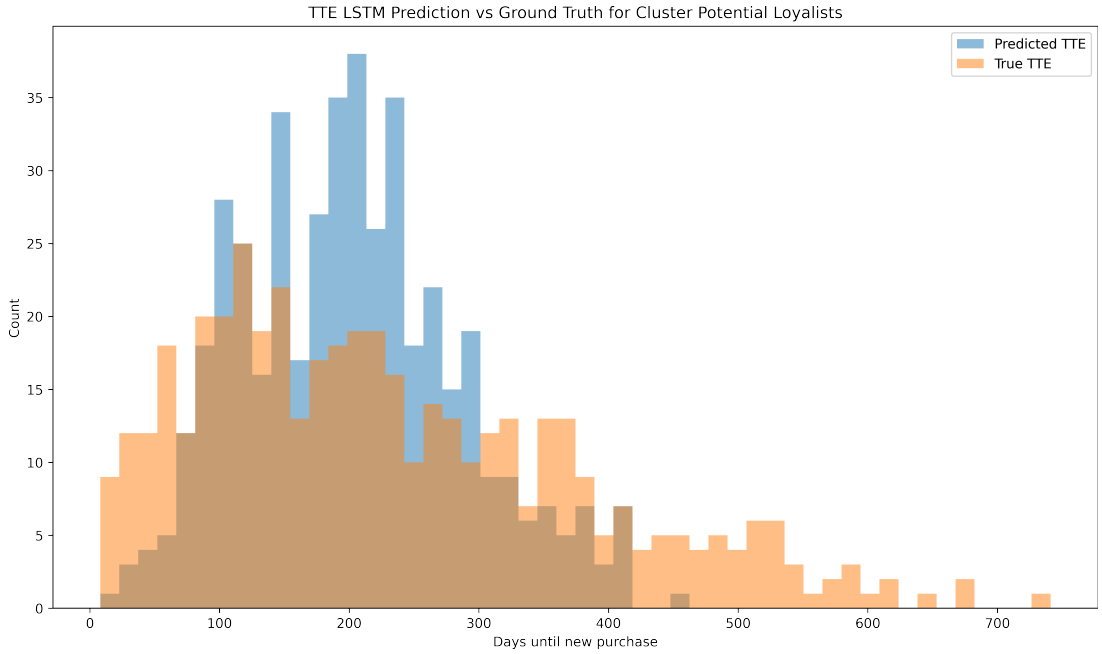
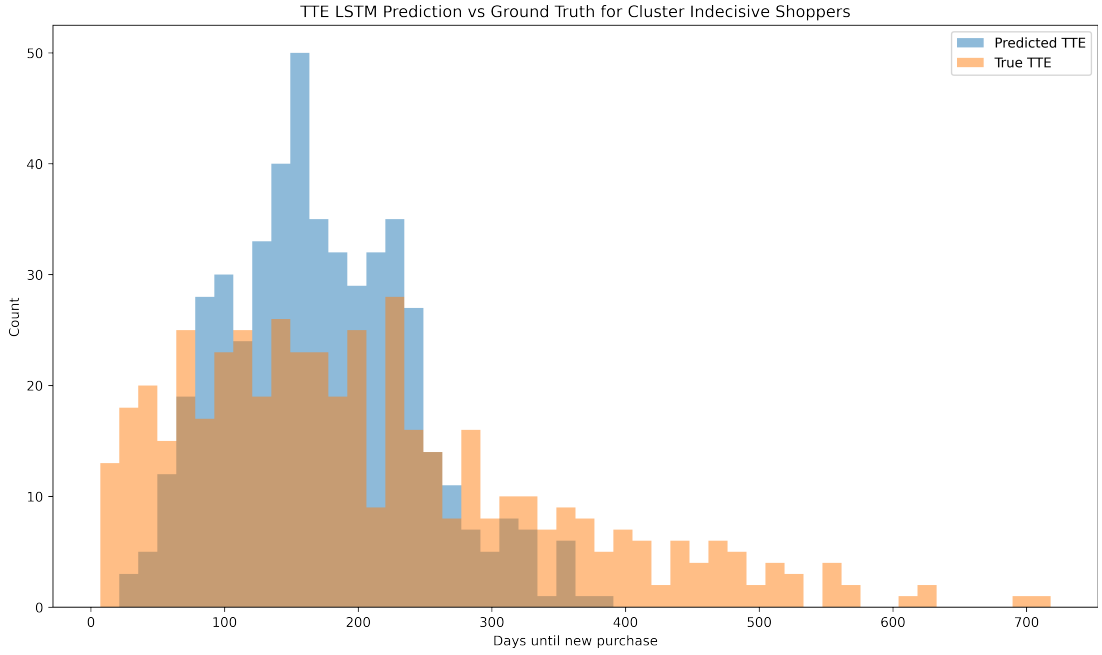


Figure 4.2: Cluster distribution along days until next purchase and average order value features.

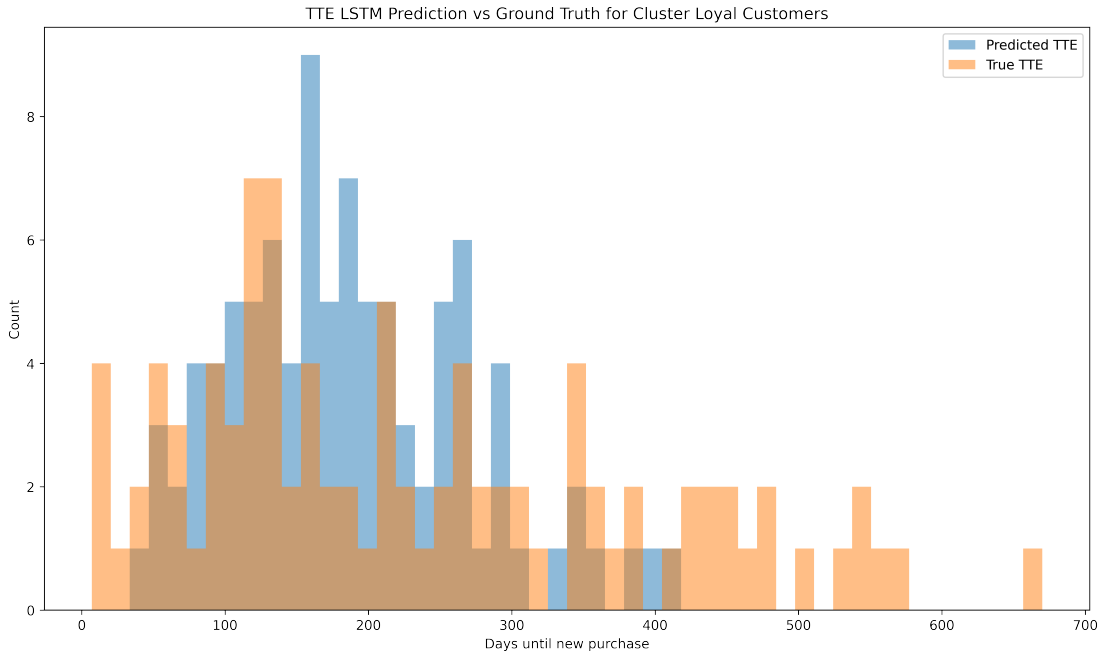


(a) TTE prediction distribution against ground truth for cluster Potential Loyalists.

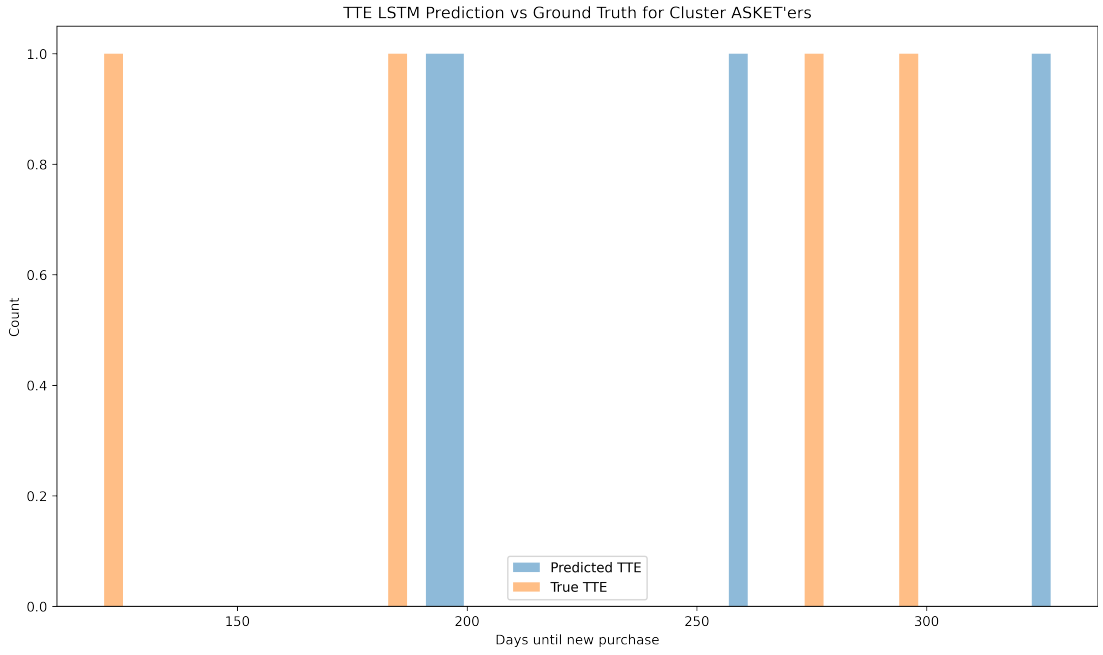


(b) TTE prediction distribution against ground truth for cluster Indecisive Customers.

Figure 4.3: Cluster-wise TTE prediction distribution against ground truth, part 1.



(a) TTE prediction distribution against ground truth for cluster Loyal Customers.



(b) TTE prediction distribution against ground truth for cluster ASKET'ers.

Figure 4.4: Cluster-wise TTE prediction distribution against ground truth, part 2.

5 | Discussion

In this chapter a discussion over the obtained results illustrated and presented in chapter 4 will follow, with particular interest addressed to the implications of all different sections of the results to the final TTE prediction and how those can be interpreted. This will hopefully prepare the ground to provide a final and coherent answer to the research question.

5.1. Clustering

The clusterization task performed on the customer dataset available for the prediction task has proven fundamental from a business point of view and an operational one to better understand the composition, behaviour and tendencies of the otherwise indistinguishable set of data points.
By the use of the elbow method, the identification of 5 clusters has been coherent with the inherent diversity of behaviour highlighted in tables 4.1 and 4.2, with a substantial enough component of customers making up each of them, except for the so called ASKET'ers, which in itself represents a group of extremely high value portion of the customers that deserve special attention also to the ends of the larger TTE predictione effort.
Further analysis and an even deeper understanding of each of these customers will prove instrumental for the company on itself to the ends of a more data-driven and context aware decision process regarding marketing, operations and customer support tasks towards the customers involved. At the same time being able to condense much of the different feature information used to obtain such a segmentation of the customer base in a single attribute is no doubt fundamental in allowing the model to obtain a better representation of each customer without the hassle of including many different attributes and the need to work in a higher dimensional representation space. It is based on such a point that the inclusion of cluster information has been regarded as a special "static" attribute in the analysis, one to be included also in the dynamic-only configuration of the dataset for two main reasons:

- It allows to easily derive graphs, statistics and detailed representations from the

results that can leverage on a well defined codification of customer habits

- the addition of a single feature has not been deleterious to the ends of maintaining a smaller representation space when working only with dynamic attributes, especially given the amount of information it inherently can provide to the models

The main takeaway from the clusterization task as such has been to better frame the contribution of static attributes to the model itself in a single attribute that on one way can provide a synthesis of all the static attributes used to derive it and at the same time is not too complex to include in the analysis with the fear that the representation space may get out of hand. At the same time though, it cannot be said that using the cluster attribute is sufficient in providing the same information to the model as all the attributes from which it is derived from, which is the reason why the second configuration mentioned in section 3.2.5 still includes both dynamic and static attributes, that is all the ones fed to the K-Prototypes model to derive the cluster belonging of each customer.

From a qualitative point of view the clusterization task clearly represents a trend provided as initial reasoning to justify such an analysis of the customers, that is the gradual shift in the company's customer base from Swedish based small buyers to a more International audience, with prominence in Germany and growing in the US, that is becoming more mature, more prone to spending, but at the same time less constant in its customer relationship with ASKET. The latter point entices the interest in conduction a TTE prediction task and largely an investigation over how to improve customer retention for the company.

5.2. TTE prediction

The TTE prediction task represents the core of the problem the thesis wants to answer and as such getting a better understanding of the obtained results can help to shed light and provide an answer to the research question.

In the context of the results obtained by a purely dynamic data configuration, what emerged has been that the model able to deal with sequential data, i.e. LSTM, was able to outperform the other ones in terms of error metrics. As already briefly mentioned in the Results section, this is mainly due to the fact that being able to work on the dynamic attributes in a sequential way allowed it to exploit its predictive edge compared to the other proposed models. This result itself is no surprise and goes to confirm the successful results that the introduction of dynamic attributes and temporal data analysis have brought to the field of churn retention and TTE prediction. This in fact represents the

current avenue of research in the field and continues to show promise in this instance as well.
It is worth stressing that much of the predictive capacity of the model, that here provides decent results without any excessive optimization, comes down to the use of RFM attributes. Such attributes are confirmed to be extremely useful when trying to analyse temporal phenomenons like purchases, subscriptions and other business interactions. Furthermore such results have been achieved in a setting that does not particularly favour DL models. Given the relative small size of the dataset and the wide variety of datapoints it presents the models cannot be overtly complex, evidenced by the relatively small amount of layers and units employed in the DL architectures in order to avoid excessive overfitting, which in general does not allow them to express their full potential. In such regard the LSTM suffers less than the MLP given that for each customer there are 12 different datapoints available, one per month, which in a sense reduces the lack of data issue.

All the previous considerations confirm that whenever a task pertains survival analysis or is framed as TTE prediction, a very good and simple enough solution to implement is the adoption of RFM attributes and a RNN structure of choice, regardless of it being a medical, legal or business application.
The predictive edge that a DL model offers is only going to improve as the amount of available data increases, but even such availability is low the possibility of increasing the length of the analysis windows and thus enriching them with more information for each customer still allow RNN to maintain an edge over traditional ML solutions.
In the context of non sequential models instead, the shortcomings of a MLP architecture are evident in the lack of data, while solutions that approach the problem in a completely different way, like tree based structures and SVM offer varying degrees of success.

Moving to the considerations regarding the second experiment, that is the TTE prediction with a static + dynamic data configuration there is the possibility to draw a definitive conclusions to the research question.
As it can be clearly observed in table 4.6 it has already been stated how the introduction of static features mainly provided an improvement of performances to the non-DL models. Compared to the previous experiments instead the LSTM struggled to keep its error down and instead suffered with such introduction. From a complexity point of view the data representation moved from a 4 dimensional space to a 20+ one, with all the consequences this brings in terms of performances and time of computation. The reason why the non-DL models improved their performance in such a setting is two-fold:

- The introduction of static attributes, which are more akin to the usual kind of data such models are built to work with, tend to be more usueful for such models instead

of a sequential based model like the LSTM

- The added complexity that DL models deal with in the representations when features increase in number is why, keeping the number of samples equal to before, MLP and LSTM performed worse. Non-DL models scale such added complexity far less and this allows them not to overfit on data as much, all the while keeping computation times low.

In light of the previous experiments, it is possible to say that the overfitting tendency of the DL models has only been worsened with the addition of many static features, that on the other hand are not even very useful to the LSTM since they remain constant at each time step in the analysis window. Despite all of this the LSTM still provides the best performance in terms of average offset and MAE, while not falling too much behind in the MSE error metric.

All of this clearly goes to show that, especially in cases such as this one where overfitting is a real problem given the small number of available samples, the addition of a significant amount of static features only goes to damage the MSE performance of the LSTM model, even by a small amount. What is interesting to see though is that despite this, the MAE does not worsen by much: what this tells us is that while the model becomes a bit worse in dealing with outliers (probably because static features regarding them are surprisingly similar to non outlier data points), its explainability of the error is not tarnished. This is confirmed by the average offsets essentially proving to be better or at least equal to before.

By observing the cluster wise behaviour in figures 4.3 and 4.4 it is clear to see that the model tends to provide a more concentrated prediction distribution compared to the real one, which is definitely to be attributed to the fact that in general customers present similar behaviours in the clusters.
Despite the single time predictions though, it may be more useful to reason in terms of the average offsets as a sort of interval where ground truths can reasonably be placed: given the value corresponding to the prediction, it makes sense to assume that with high probability the customer, if not placing the purchase exactly at the predicted time, will instead do so in an interval of [*Predicted time - Average negative offset; Predicted time + Average positive offset*], which could also be further refined by including the cluster average negative and positive offsets instead of the general one, given how extensive the correlation in behaviours between clusters is.
Such a solution, which can definitely take on different meaning given the use case, may

be useful in defining what essentially represents a "period of interest" for each customer: during such period an event is likely to happen or expected, thus it can be useful in a business case for example to try and engage said customer more towards the end in case no purchase gets placed, in order to ward off the chance of it churning. At the same time, whenever it is expected from the prediction that the purchase will be placed after what is generally a more common behaviour in the cluster, this can itself be interpreted as a sign that the customer may be losing interest in the company and eagerness in buying. Finally, directly referencing in such case the interpretation framework, the most important "alarm signal" is represented by an event being predicted to happen before the date the analysis is run, which in a e-commerce scenario means churn and could mean that there is much interest in expending resources trying to keep the customer hooked, but instead more of a re-activation tactic should be employed in interactions and communications towards them.

6 | Conclusions

In this chapter a summary of the whole thesis with specific reference to its finding and contributions will follow, trying to provide a critical view on the obtained results.

6.1. Contributions

This work focuses on the task of applying survival analysis to a specific application, churn prediction, by analysing the contribution to reducing performance errors when using ML models.
By observing how the addition of static features to a dataset consisting of dynamic ones, derived from the RFM framework, contributes to the error performances of models has shed light on how to optimise such a task and makes a case on whether or not such additions make sense in order to continue improving on the current State of the Art.
The thesis also offers a simple, yet effective framework of interpretation that neatly fits in the gap existing between traditional churn prediction tasks and TTE prediction in section B.
As a final point the thesis provides a sound and complete overview of how to achieve a more context aware view of the problem, by employing a clusterization task to better describe the different behaviours of customer cohorts. This has proven to be very useful to the ends of results interpretability and goes to show that when such an approach is soundly conducted it can really change the shape in which conclusions can be drawn on a task.

6.2. Answer to the Research Question

After all the previous considerations we can finally provide an answer to the Research Question of this thesis:

- Do static features improve the predictive power by reducing the error (MSE) in time-to-event prediction problems?

In the specific application of TTE prediction problems the contribution provided by static features can be seen as a positive one, being that the enriched contextual representation that it provides to the data tends to provide equal if not smaller error measurements when other operational factors hindering the performances are not present.
In particular reference to the kind of analysed architectures:

- Non-DL models showed reduced error measurements when static features were included in the dataset in all cases. Such a contribution is considerable and especially in cases where data scarcity is a characteristic of the task, can bring such models' performances on par with more powerful architectures, without the added computational load
- DL models do not generally suffer from an increase in error measurements when static features complement the datasets, provided that the issue of overfitting is not present. In more general terms the addition of further features in the dataset makes the representation space more complex and given the "data-hungry" nature of DL models, can actually get in the way of performances by causing the models to overfit. When this is not the case instead, the models experience a slight improvement in error reduction, although it is debatable that the static nature of the data is actually a benefit or not in such a case.

6.3. Validity of the results

Based on the limitations that along the thesis have added themselves to the work and the assumptions of the research work, it is necessary to put out some justifications on how reliable the answers provided are. Central to the issue is that the project has been conducted with limited resources and that based on this, the set of results obtained from experiments feature a limited amount of samples, customer order transactions while business, industrial or medical settings may have different requirements than what is provided. Despite this, the techniques that are employed are all general purpose and there is room to generalise the answer to the research question to other application cases.

In particular the limitations on the composition dataset is according to what makes sense from a business point of view based on the host company's, ASKET's, requirements and this may of course not prove general enough. At the same time some other operational limitations that have been put in place directly reference the kind of interpretation and theoretical framework laid out for the task, such as for example the different ways datasets have had to be configured in order to accommodate for sequential and non ML models. The intention all along has been to try and make such difference as little as possible in

order not to impact on the final outcome, but at the same time there are others that are inherent to the models and are unavoidable.

The specific use case also allowed for what could be a much richer contextual awareness than usual, especially in the capability of extracting from the dataset clusters of immediate and reasonable interpretation that definitely made the task of assessing model behaviour among them much easier. This is obviously not always true and as such, it makes sense that not all applications of this kind may end up featuring such a step in the analysis.

Finally, while the choice of models to be used in the comparison for TTE prediction has been made in order to account for the most common ones used in the literature, such representative models cannot possibly account for considerations spanning the whole ML world. The limitation has mainly been dictated by how a model offers itself to a regression task and how wide spread its utilization has been in survival analysis and churn retention, where more often than newer architectures struggle to be employed.

6.4. Future work

As the field of survival analysis continues progressing, avenues of research to be explored are confirmed to be in the footprints of what has been laid out in the work of [66]. Using regression formulations of the problem that can leverage on sequential structures is definitely one of the most promising ways to continue researching.

What can be gathered from the experience of this thesis is that expanding on its findings in an application where the lack of data does not hinder DL model performances, it can be further shown that static features have a positive effect on reducing the error. This research in itself is twofold: on the one hand improving on RNN architectures in order to allow them to handle more features without the risk of overfitting can push their performances to the best achievable. On the other hand instead such models are very complex, power and data hungry and maybe not always suited to applications in smaller studies where data may be a problem. In this case non-DL architectures are to be preferred and possibly expanded upon in order to make them able to perform equally if not better than such more complex ones.

Another very important avenue for research is represented by continuing the shift of churn retention problems from classification tasks, which are quite limited, requiring the imposition of strict limitations and targets while not conveying a global view of the results, to more of a regression problem. In this latter case, obtaining distributions over the data can effectively be more informative than just knowing that in a specific setting an event

will happen or not. Instead they would allow to analyse the problem with more degrees of freedom, which would greatly benefit the decision making progress in critical applications, like in medical monitorings where the presence of an event under an umbrella of possible conditions could spell the difference between a good assessment and effective cure to an ill performed diagnosis.

Bibliography

[1] *Strategic database marketing*. 1994.

[2] 03 2011. URL https://silo.pub/mathematical-statistics-basic-ideas-and-selected-html.

[3] 05 2020. URL https://www.returnlogic.com/blog/what-is-customer-churn-in-ecommerce/.

[4] 7 weird and wonderful applications of artificial intelligence, 05 2021. URL https://www.cityam.com/7-weird-and-wonderful-applications-of-artificial-intelligence/.

[5] 99Content. 44 fascinating ecommerce statistics - 2022 update, 05 2021. URL https://99firms.com/blog/ecommerce-statistics.

[6] N. Alboukaey, A. Joukhadar, and N. Ghneim. Dynamic behavior based churn prediction in mobile telecom. *Expert Systems with Applications*, 162:113779, 2020. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113779. URL https://www.sciencedirect.com/science/article/pii/S0957417420306035.

[7] Amazon. Small businesses, 09 2020. URL https://www.aboutamazon.com/impact/empowerment/small-businesses.

[8] M. Arif Wani, M. Kantardzic, and M. Sayed-Mouchaweh. *Trends in Deep Learning Applications*, page 1–7. Advances in Intelligent Systems and Computing. Springer, 2020. ISBN 9789811518164. doi: 10.1007/978-981-15-1816-4_1. URL https://doi.org/10.1007/978-981-15-1816-4_1.

[9] A. Banerjee and H. Shan. *Model-Based Clustering*, pages 686–689. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_554. URL https://doi.org/10.1007/978-0-387-30164-8_554.

[10] T. W. Bank. World bank sme finance: Development news, research, data, 2021. URL https://www.worldbank.org/en/topic/smefinance.

[11] D. Bell and C. Mgbemena. Data-driven agent-based exploration of customer behavior. *SIMULATION*, 94(3):195–212, 2018. doi: 10.1177/0037549717743106. URL https://doi.org/10.1177/0037549717743106.

[12] R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, 2008. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2006.11.006. URL https://www.sciencedirect.com/science/article/pii/S1386505606002747.

[13] D. S. Bhat, K. Kansana, and J. Majid. A review paper on e-commerce. 02 2016.

[14] BigCommerce. Top 14 ecommerce trends in 2021 (+ industry experts' insight), 02 2022. URL https://www.bigcommerce.com/articles/ecommerce/ecommerce-trends/.

[15] J. Bult and T. Wansbeek. Optimal selection for direct mail. *Marketing Science*, 14: 378–394, 11 1995. doi: 10.1287/mksc.14.4.378.

[16] M. A. Camilleri. The use of data-driven technologies for customer-centric marketing. *International Journal of Big Data Management*, 1, 01 2019. doi: 10.1504/IJBDM.2019.10023294.

[17] A. Castro-López, V. Iglesias, and J. Puente. Slow fashion trends: Are consumers willing to change their shopping behavior to become more sustainable? *Sustainability*, 13(24), 2021. ISSN 2071-1050. doi: 10.3390/su132413858. URL https://www.mdpi.com/2071-1050/13/24/13858.

[18] M. Chen. The fast rise of slow fashion - jumpstart magazine, 06 2020. URL https://www.jumpstartmag.com/the-fast-rise-of-slow-fashion/.

[19] J. M. Chua. The environment and economy are paying the price for fast fashion — but there's hope, 09 2019. URL https://www.vox.com/2019/9/12/20860620/fast-fashion-zara-hm-forever-21-boohoo-environment-cost.

[20] H. Clark. Slow fashion an oxymoron or a promise for the future? *Fashion theory*, 12(4):427–446, 2008.

[21] E. Cramer-Flood. Global ecommerce forecast 2022, 02 2022. URL https://www.emarketer.com/content/global-ecommerce-forecast-2022.

[22] S. De, P. P, and J. Paulose. Effective ml techniques to predict customer churn. pages 895–902, 09 2021. doi: 10.1109/ICIRCA51532.2021.9544785.

[23] A. De Caigny, K. Coussement, and K. W. De Bock. A new hybrid classifi-

cation algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2):760–772, 2018. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.02.009. URL https://www.sciencedirect.com/science/article/pii/S0377221718301243.

[24] E. Domingos, B. Ojeme, and O. Daramola. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. *Computation*, 9(3), 2021. ISSN 2079-3197. doi: 10.3390/computation9030034. URL https://www.mdpi.com/2079-3197/9/3/34.

[25] F. R. Dwyer. Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing*, 11(4):6–13, 1997. doi: https://doi.org/10.1002/(SICI)1522-7138(199723)11:4<6::AID-DIR3>3.0.CO;2-T. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291522-7138%28199723%2911%3A4%3C6%3A%3AAID-DIR3%3E3.0.CO%3B2-T.

[26] Encircled. Circular fashion at encircled. URL https://www.encircled.ca/pages/end-of-life-cycle.

[27] C. Esteban, D. Schmidt, D. Krompaß, and V. Tresp. Predicting sequences of clinical events by using a personalized temporal latent embedding model. In *2015 International Conference on Healthcare Informatics*, pages 130–139, 2015. doi: 10.1109/ICHI.2015.23.

[28] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. pages 93–101, 10 2016. doi: 10.1109/ICHI.2016.16.

[29] P. Fader, B. Hardie, Y. Liu, J. Davin, and T. Steenburgh. 'how to project customer retention' revisited: The role of duration dependence. *SSRN Electronic Journal*, 43, 01 2018. doi: 10.2139/ssrn.3100346.

[30] K. Fletcher. Slow fashion: An invitation for systems change. *Fashion Practice*, 2 (2):259–265, 2010. doi: 10.2752/175693810X12774625387594. URL https://doi.org/10.2752/175693810X12774625387594.

[31] R. Flynn. Survival analysis. *Journal of Clinical Nursing*, 21(19–20):2789–2797, 10 2012. ISSN 1365-2702. doi: 10.1111/j.1365-2702.2011.04023.x.

[32] C. Franco and B. S. Advantages and challenges of e-commerce customers and businesses: in indian perspective. *International Journal of Research*, 4:7–13, 03 2016. doi: 10.29121/granthaalayah.v4.i3SE.2016.2771.

[33] K. Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[34] A. Gallo. The value of keeping the right customers. *Harvard Business Review*, 10 2014. ISSN 0017-8012. URL https://hbr.org/2014/10/the-value-of-keeping-the-right-customers.

[35] M. K. Goel, P. Khanna, and J. Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.

[36] A. Granov. *Customer loyalty, return and churn prediction through machine learning methods: for a Swedish fashion and e-commerce company*. PhD thesis, 2021. URL http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-184709.

[37] GuruFocus. Uk online shopping and e-commerce statistics for 2017, 2022. URL https://www.nasdaq.com/articles/uk-online-shopping-and-e-commerce-statistics-2017-2017-03-14.

[38] J. Hadden, A. Tiwari, R. Roy, and D. Ruta. Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10):2902–2917, 2007. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2005.11.007. URL https://www.sciencedirect.com/science/article/pii/S0305054805003503.

[39] O. Heinisch. Steel, r. g. d., and j. h. torrie: Principles and procedures of statistics. (with special reference to the biological sciences.) mcgraw-hill book company, new york, toronto, london 1960, 481 s., 15 abb.; 81 s 6 d. *Biometrische Zeitschrift*, 4: 207–208, 1962.

[40] X.-x. Huang, Z.-p. Hu, C.-s. Liu, D.-j. Yu, and L.-f. Yu. The relationships between regulatory and customer pressure, green organizational responses, and green innovation performance. *Journal of Cleaner Production*, 112:3423–3433, 2016.

[41] Z. Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34. Citeseer, 1997.

[42] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 09 1998. ISSN 1573-756X. doi: 10.1023/A:1009769707641. URL https://doi.org/10.1023/A:1009769707641.

[43] E. J. Humphrey. An exploration of deep learning in content-based mu-

sic informatics - proquest, 2015. URL https://www.proquest.com/openview/50311b3db34eed99649482493e46e3a9/1?pq-origsite=gscholar&cbl=18750.

[44] E. J. Humphrey, J. P. Bello, and Y. LeCun. Feature learning and deep architectures: new directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 12 2013. URL https://www.proquest.com/scholarly-journals/feature-learning-deep-architectures-new/docview/1475519729/se-2?accountid=28385. Date revised - 2014-01-01; Number of references - 1; Last updated - 2014-02-14.

[45] C.-L. Hwang and K. Yoon. *Methods for Multiple Attribute Decision Making*, page 58–191. Lecture Notes in Economics and Mathematical Systems. Springer, 1981. ISBN 9783642483189. doi: 10.1007/978-3-642-48318-9_3. URL https://doi.org/10.1007/978-3-642-48318-9_3.

[46] A. Håkansson. Portal of research methods and methodologies for research projects and degree projects. In *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering FECS'13*, page 67–73. CSREA Press U.S.A, 2013. ISBN 1-60132-243-7. URL http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-136960. [ed] Hamid R. Arabnia Azita Bahrami Victor A. Clincy Leonidas Deligiannidis George Jandieri.

[47] incify. 70 useful e-commerce statistics for 2022, 02 2022. URL https://www.incify.co/70-useful-e-commerce-statistics-for-2022/.

[48] B. Institute. 46 cart abandonment rate statistics – cart & checkout – baymard institute, 2022. URL https://baymard.com/lists/cart-abandonment-rate#why-users-abandon-their-cart.

[49] Invesp. Customer acquisition statistics and trends, 2022. URL https://visual.ly/community/Infographics/business/customer-acquisition-statistics-and-trends.

[50] N. Jain, A. Tomar, and P. K. Jana. A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. *Journal of Intelligent Information Systems*, 56(2):279–302, 04 2021. ISSN 1573-7675. doi: 10.1007/s10844-020-00614-9. URL https://doi.org/10.1007/s10844-020-00614-9.

[51] L. Jimenez and D. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(1):39–54, 1998. doi: 10.1109/5326.661089.

[52] H. Jing and A. J. Smola. Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 515–524, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346757. doi: 10.1145/3018661.3018719. URL https://doi.org/10.1145/3018661.3018719.

[53] S. Jung and B. Jin. A theoretical investigation of slow fashion: sustainable future of the apparel industry. *International Journal of Consumer Studies*, 38(5):510–519, 2014. doi: https://doi.org/10.1111/ijcs.12127. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ijcs.12127.

[54] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh. Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. volume 3, 01 2010. doi: 10.1016/j.procs.2010.12.011.

[55] S. Khan and T. Yairi. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107:241–265, 2018. ISSN 0888-3270. doi: https://doi.org/10.1016/j.ymssp.2017.11.024. URL https://www.sciencedirect.com/science/article/pii/S0888327017306064.

[56] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer New York, 12 2010. ISBN 9781441929853.

[57] D. Koletsi and N. Pandis. Survival analysis, part 2: Kaplan-meier method and the log-rank test. *American Journal of Orthodontics and Dentofacial Orthopedics: Official Publication of the American Association of Orthodontists, Its Constituent Societies, and the American Board of Orthodontics*, 152(4):569–571, 10 2017. ISSN 1097-6752. doi: 10.1016/j.ajodo.2017.07.008.

[58] I. Kononenko. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7(4):317–337, 1993. doi: 10.1080/08839519308949993. URL https://doi.org/10.1080/08839519308949993.

[59] KPMG. The truth about online customers, 2022. URL https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/01/the-truth-about-online-consumers.pdf.

[60] S. Lamrhari, H. E. Ghazi, M. Oubrich, and A. E. Faker. A social crm analytic framework for improving customer retention, acquisition, and conversion. *Technological Forecasting and Social Change*, 174:121275, 2022. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2021.121275. URL https://www.sciencedirect.com/science/article/pii/S0040162521007095.

[61] A. Langdown. Slow fashion as an alternative to mass production: A fashion practitioner's journey. *Social Business*, 4(1):33–43, 2014.

[62] H. Li. Deep learning for natural language processing: advantages and challenges. *National Science Review*, 5(1):24–26, 09 2017. ISSN 2095-5138. doi: 10.1093/nsr/nwx110. URL https://doi.org/10.1093/nsr/nwx110.

[63] H.-M. Lin, J. M. Williamson, and H.-Y. Kim. Firth adjustment for weibull current-status survival analysis. *Communications in Statistics - Theory and Methods*, 49 (18):4587–4602, 2020. doi: 10.1080/03610926.2019.1606241. URL https://doi.org/10.1080/03610926.2019.1606241.

[64] T. B. R. P. LTD. Sustainable fashion market analysis shows the market progress in attempt to decrease pollution in the global ethicalfashion market 2020, 10 2020. URL https://www.globenewswire.com/news-release/2020/10/28/2116073/0/en/Sustainable-Fashion-Market-Analysis-Shows-The-Market-Progress-In-Attempt-To-De html.

[65] C. Marquis. What does slow fashion 'actually' mean? URL https://www.forbes.com/sites/christophermarquis/2021/05/14/what-does-slow-fashion-actually-mean/.

[66] E. Martinsson. Wtte-rnn - less hacky churn prediction · focus on the objective, 12 2016. URL https://ragulpr.github.io/2016/12/22/WTTE-RNN-Hackless-churn-modeling/.

[67] E. Martinsson. Wtte-rnn: Weibull time to event recurrent neural network a model for sequential prediction of time-to-event in the case of discrete or continuous censored data, recurrent events or time-varying covariates. 2017. URL https://odr.chalmers.se/handle/20.500.12380/253611.

[68] R. Matthew McCutchen and S. Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In A. Goel, K. Jansen, J. D. P. Rolim, and R. Rubinfeld, editors, *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 165–178, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-85363-3.

[69] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller. *Concrete Problems for Autonomous Vehicle Safety: Advantages of Bayesian Deep Learning*. International Joint Conferences on Artificial Intelligence, Inc., 08 2017. ISBN 9780999241103. doi: 10.17863/CAM.12760. URL https://www.repository.cam.ac.uk/handle/1810/266683.

[70] P. Mechanics. 65 best inventions of the past 65 years. *Popular Mechanics*, 2018.

[71] C. G. Mena, A. De Caigny, K. Coussement, K. W. De Bock, and S. Lessmann. Churn prediction with sequential data and deep neural networks. a comparative analysis, 2019. URL https://arxiv.org/abs/1909.11114.

[72] T. Mutanen. Customer churn analysis–a case study. *Journal of Product and Brand Management*, 14(1):4–13, 2006.

[73] G. Nathiya, S. Punitha, and M. Punithavalli. An analytical study on behavior of clusters using k means, em and k* means algorithm. *arXiv preprint arXiv:1004.1743*, 2010.

[74] D. Neel. how is cac changing over time?, 2021. URL https://www.profitwell.com/recur/all/how-is-cac-changing-over-time.

[75] L. L. Nguyen and R. E. Scully. *Chapter 1 - Epidemiology and Research Methodology*, volume 1, pages 1–12. Elsevier inc., ninth edition edition, 2019. ISBN 9780323775571. URL https://www.clinicalkey.com/#!/content/book/3-s2.0-B9780323427913000013.

[76] N. D. Nguyen, T. Nguyen, and S. Nahavandi. System design perspective for human-level agents using deep reinforcement learning: A survey. *IEEE Access*, PP:1–1, 11 2017. doi: 10.1109/ACCESS.2017.2777827.

[77] A. Perišić, D. Šišak Jung, and M. Pahor. Churn in the mobile gaming field: Establishing churn definitions and measuring classification similarities. *Expert Systems with Applications*, 191:116277, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.116277. URL https://www.sciencedirect.com/science/article/pii/S0957417421015852.

[78] O. Pinnock. Sustainable fashion wants brands to redefine business growth, 09 2021. URL https://www.forbes.com/sites/oliviapinnock/2021/09/24/degrowth-is-trending-in-sustainable-fashion-what-does-that-mean-for-brands/.

[79] N. Remy, E. Speelman, and S. Swartz. Style that's sustainable: A new fast-fashion formula | mckinsey, 10 2016. URL https://www.mckinsey.com/business-functions/sustainability/our-insights/style-thats-sustainable-a-new-fast-fashion-formula.

[80] A. Rotar. ecommerce report 2021 - fashion. RAND Corp., Santa Monica, CA, USA, RR-1776-NYCEDC, 07 2021. URL https://www.statista.com/study/38340/ecommerce-report-fashion/.

[81] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. ISSN 0306-4573. doi: https://doi.org/10.1016/0306-4573(88)90021-0. URL https://www.sciencedirect.com/science/article/pii/0306457388900210.

[82] J. Santora. 5 cart abandonment stats to help you win "lost" sales now, 11 2020. URL https://optinmonster.com/cart-abandonment-statistics/.

[83] P. Schober and T. R. Vetter. Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia and Analgesia*, 127(3):792–798, 09 2018. ISSN 0003-2999. doi: 10.1213/ANE.0000000000003653. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6110618/.

[84] C. Shen, D. Nguyen, Z. Zhou, S. B. Jiang, B. Dong, and X. Jia. An introduction to deep learning in medical physics: advantages, potential, and challenges. *Physics in Medicine & Biology*, 65(5):05TR01, 03 2020. doi: 10.1088/1361-6560/ab6f51. URL https://doi.org/10.1088/1361-6560/ab6f51.

[85] M. Stepanova and L. Thomas. Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289, 2002. doi: 10.1287/opre.50.2.277.426. URL https://doi.org/10.1287/opre.50.2.277.426.

[86] G. Taher. E-commerce: Advantages and limitations. volume 11, pages 153–165. hrmars, 02 2021.

[87] D. J. F. M. Thuijs, G. L. Hickey, and R. L. J. Osnabrugge. Statistical primer: basics of survival analysis for the cardiothoracic surgeon. *Interactive Cardiovascular and Thoracic Surgery*, 27(1):1–4, 07 2018. ISSN 1569-9285. doi: 10.1093/icvts/ivy010.

[88] H. Tolley, J. Barnes, and M. Freeman. Chapter 10 - survival analysis. In M. D. Freeman and M. P. Zeegers, editors, *Forensic Epidemiology*, pages 261–284. Academic Press, Amsterdam, 2016. ISBN 978-0-12-404584-2. doi: https://doi.org/10.1016/B978-0-12-404584-2.00010-0. URL https://www.sciencedirect.com/science/article/pii/B9780124045842000100.

[89] D. Tolstoy, E. R. Nordman, S. M. Hånell, and N. Özbek. The development of international e-commerce in retail smes: An effectuation perspective. *Journal of World Business*, 56(3):101165, 2021. ISSN 1090-9516. doi: https://doi.org/10.1016/j.jwb.2020.101165. URL https://www.sciencedirect.com/science/article/pii/S1090951620300936.

[90] C.-F. Tsai and Y.-H. Lu. Customer churn prediction by hybrid neural networks. *Expert Syst. Appl.*, 36:12547–12553, 12 2009. doi: 10.1016/j.eswa.2009.05.032.

[91] Unctad. Global e-commerce jumps to $26.7 trillion, 05 2021. URL `https://unctad.org/news/global-e-commerce-jumps-267-trillion-covid-19-boosts-online-sales`.

[92] L. Van Der Maaten, E. Postma, J. Van den Herik, et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

[93] P. C. Van Dijk, K. J. Jager, A. H. Zwinderman, C. Zoccali, and F. W. Dekker. The analysis of survival data in nephrology: basic concepts and methods of cox regression. *Kidney international*, 74(6):705–709, 2008.

[94] L. Wang, H. Wang, Y. Huang, B. Yan, Z. Chang, Z. Liu, M. Zhao, L. Cui, J. Song, and F. Li. Trends in the application of deep learning networks in medical image analysis: Evolution between 2012 and 2020. *European Journal of Radiology*, 146:110069, 2022. ISSN 0720-048X. doi: https://doi.org/10.1016/j.ejrad.2021.110069. URL `https://www.sciencedirect.com/science/article/pii/S0720048X21005507`.

[95] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. 51(6), 02 2019. ISSN 0360-0300. doi: 10.1145/3214306. URL `https://doi.org/10.1145/3214306`.

[96] X. Wen, Y. Wang, X. Ji, and M. K. Traoré. Three-stage churn management framework based on dcn with asymmetric loss. *Expert Systems with Applications*, 207:117998, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117998. URL `https://www.sciencedirect.com/science/article/pii/S0957417422012222`.

[97] C. J. Willmott and K. Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82, 12 2005. ISSN 0936-577X, 1616-1572. doi: 10.3354/cr030079. URL `https://www.int-res.com/abstracts/cr/v30/n1/p79-82/`.

[98] Y. Xie, X. Li, E. Ngai, and W. Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3, Part 1):5445–5449, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.06.121. URL `https://www.sciencedirect.com/science/article/pii/S0957417408004326`.

[99] G. Yang, Y. Cai, and C. K. Reddy. Spatio-temporal check-in time prediction with recurrent neural network based survival analysis. In *Proceedings of the Twenty-*

Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 2976–2983. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/413. URL https://doi.org/10.24963/ijcai.2018/413.

[100] K. D. Young, J. J. Menegazzi, and R. J. Lewis. Statistical methodology: Ix. survival analysis. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 6(3):244–249, 03 1999. ISSN 1069-6563. doi: 10.1111/j.1553-2712.1999.tb00165.x.

[101] B. Zupan, J. Demšar, M. W. Kattan, J. Beck, and I. Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, 20(1):59–75, 2000. ISSN 0933-3657. doi: https://doi.org/10.1016/S0933-3657(00)00053-1. URL https://www.sciencedirect.com/science/article/pii/S0933365700000531. Selected Papers from AIMDM '99.

Appendix - Contents

A | Extended background

A.1. Previous works

In this section deeper analyses of referenced works in the field of survival analysis, churn retention and TTE prediction will be provided.

Let's analyse now the characteristics of many notable studies in the field of survival analysis, which are based on the application of ML techniques together with context-related knowledge.

First off it is worth starting with a paper clearly showcasing the main characteristics of the analysis of churn prediction in transactional datasets. [54] applies RFM analysis and an extended RFM with an additional feature representing the total number of items purchased by a customer in addition to the number of orders. The models have been used to predict customer purchase behaviour and estimate Customer Lifetime Value (CLV). [15] defined RFM as:

- R (Recency): the period since the last purchase; a lower value corresponds to a higher probability of the customers making a repeated purchase.
- F (Frequency): number of purchases made within a certain period; higher frequency indicates greater loyalty.
- M (Monetary): the money spent during a certain period; a higher value indicates that the company should focus more on that customer.

The prototype of a very loyal customer thus presents low recency, high frequency and high monetary value. Given that in some alternative models the three parameters are considered with different relevance according to the specific application, the researchers found that the regular and extended RFM performed similarly in the task of using the K-means algorithm to highlight similar clusters of customers. Through business expertise a final CLV estimation was derived. Approaches like this one enable us to get a better understanding on how to tackle survival analysis problems and in this particular instance churn prediction. In such a case it makes sense to organise the inputs in a way that

emphasises the intrinsic time dependency of the data and its evolution over time. In order to accommodate for such handling of the data, one necessarily has to proceed with discretisation techniques like averaging and extraction of representative values in those models that lack a built-in refined understanding of historical data that keeps on progressing and evolving, like in RNN and LSTM models.

[90] analysed the issue of customer churn through the scope of DL techniques: with the goal of developing a two step approach involving an initial discarding of unrepresentative training data and then the development of a prediction model, they resorted to the employment of two distinct techniques: a supervised prediction model through Back-Propagation ANNs and an unsupervised clustering model through SOM. Such models were combined and their results analysed, highlighting the better performance provided by using ANN in both of the aforementioned tasks.

As evidenced by the comprehensive review on the current ML techniques used in customer churn prediction by [22], the current emphasis has been on experimenting with hybrid and ensemble techniques, as well as social network analysis as a very useful tool to predict customer churn in telecom. On this note, it is worthy to analyse the developments provided by some of the reported papers.
[71] analyses the problem of churn retention in a telecom setting and stresses the value of time varying attributes compared to static ones, e.g. demographic indicators. Through the comparison of the attributes provided by a RFM framework, the authors explain how off-the-shelf models like logistic regression are very rarely able to deal with time varying attributes and instead need to aggregate them in order to make them useful to the model. Such a modification in the data actually proves to be deleterious towards the performance of the model, thus the employment of RNN models that are in fact able to leverage the temporal nature of the data manages to improve the performances significantly. The comparison mainly relates to off-the-shelf models and by now it can still provide interesting insights when performed on more recent models, like ensemble ones. Furthermore it emphasises a very relevant point: most of the current datasets available for survival analysis either come from medical or business settings, with the latter usually benefitting from a wide availability of RFM data and generally better reflecting the nature of the data, that comes with time dependent attributes and censored data.
[6] follows on a similar note in the analysis of the customer churn problem in the telecommunication sector. This time the focus is put on the scope of the analysis. The author emphasises the dynamism of such a sector that requires faster and more efficient predictions compared to the current practice of performing monthly predictions without looking at trends that instead appear when detailing the analysis to a daily scope. Furthermore

the paper provides a comparison of two categories of models: statistical-based ones and DL ones. In this respect, the results clearly show that the latter category performs much better than the first one because of the richer underlying representation that DL architectures get from data. It is further shown that the employment of RFM attributes in the analysis enhance the results and make both the DL best performing models on the same level in terms of performance. This clearly goes to show the prominence of deeper networks when managing to exploit time varying data.

[50] proposes a new approach for analysing employee churn involving MADM paired up with ML techniques, CatBoost in particular. As the authors refer to it, the Employee Churn Prediction and Retention (ECPR). The ECPR is based on Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [45]. The authors first formulate a novel accomplishment-based employee importance model (AEIM) in order to group the employees into three categories based on the accomplishment parameters. The class-wise employee churn using the CatBoost algorithm is then predicted and compared with the results of other state-of-the-art ML algorithms. The authors finally provide a class-wise employee retention strategy using a permutation-based feature importance method leveraging on prediction results.

[23] proposes a new hybrid classification method for customer churn prediction that builds upon previous experimentation performed through DT and Logistic Regression (LR) and tries to combine it into a new model called LLM. LLM scores significantly better than its building blocks logistic regression and decision trees and performs at least as well as more advanced ensemble methods random forests and logistic model trees. Comprehensibility is addressed by a case study for which benefits are observed by using the LLM compared to using decision trees or logistic regression. Further expansion on the model can be enacted by employing a SVM instead of LR in the model, but it poses as a really good advancement in the literature since typically the choice for a classification technique is a trade-off between predictive performance and comprehensibility and the LLM scores high on both requirements.

[29] conducted an analysis of the Beta-geometric distribution in its application to cohort-level retention rates in order to build a more robust model and improve its predictions. Given the notion that increasing cohort-level retention rates are purely due to cross-sectional heterogeneity, an individual customer's propensity to churn supposedly does not change over time. In order to study this assumption the author propose a modification of the model, the beta-discrete-Weibull. It allows individual-level churn probabilities to increase or decrease over time. The most relevant aspect of the finding is that even

when aggregate retention rates are monotonically increasing, the individual-level churn probabilities are unlikely to be declining over time. Nonetheless the relevance of such trends in the model do not offset the value of cross-sectional heterogeneity, thus confirming the validity and usefulness of the Beta-geometric distribution.

Some very promising findings in the matter of defining how to model churn, are showcased by [66], where the author, rather than focusing on "the absence of an event", decides to shift the attention towards the concept of TTE, that is literally the amount of time until the next event (e.g. a purchase). This TTE could potentially tend to infinity, indicating churn, or at least give an estimate of the Remaining Useful Life (RUL) of a customer. In his approach the author uses a specific model introduced in [67], the so called WTTE-RNN, a recurrent neural network leveraging on the Weibull distribution to perform predictions on the TTE. The event is thought of under the influence of survival analysis and can thus be interpreted as the moment a machine becomes unusable as well as the next purchase in time, like in a fashion application case.

A.2. History and rise of survival analysis

Survival analysis was developed to assess patient survival, and while death is often the primary event of interest, survival analysis can also be used to assess treatment failure, such as time to loss of graft patency or amputation. Rather than simply addressing frequency, survival analysis also captures an element of time to an event. It also incorporates censorship, in which data about the event of interest are unknown because of withdrawal of the patient from the study [75]. Further developments were also obtained in related cases bound to legal proceedings and more specifically determination of loss or damage [88] as well as completely different fields, mainly pertaining loan practices, credit ratings and assessments of risks in insurances [85]. There are various applications drawing from the results of such a field, like for example estimations of failure of mechanical components, but it can be applied to anything that provides observations of measured time until a given event, in particular referring to trying to apply its results and possible insights to more commercial applications. Examples of those are the estimation of length of contract in subscription-based services or in any kind of context in which a more or less defined and rigid relationship between customer and company is defined.
From its early beginnings in the 1980s, survival analysis has been the reason for development of many different statistical techniques, such as various forms of estimators, like the Kaplan-Meyer and Cox Proportional Methods that are further referenced and analysed in 2.1. Developments in the field of ML have also found an optimal avenue for application

in survival analysis problems. We can observe most of these results in the works related to the medical field, but in later years business and commercial applications have also gained a lot of traction, starting to provide useful insights thanks to the exponentially increasing availability of data they can boast. [101] provides one of the earliest examples of applications of ML to the medical field. In its analysis on the recurrence of prostate cancer they try to deal with the difficulties of handling censored data with regular ML models. In their result they highlight a solution that makes any ML algorithm easily portable to survival analysis, solving issues related with censored data.

Stemming from many similar works like [101] ML models have thus begun to be deeply integrated in many survival analysis solutions, with many accounts of the validity and usefulness of employing data-centric techniques in commercial settings. One of the most prominent of those is churn retention, which in later years has gathered a lot of attention and continues to be a focus in research especially in subscription or contract-based systems, as shown in [11]. On a similar note, [60] introduces new analytical inferences based on social network information from customers in order to improve existing Customer Relationship Management (CRM) systems and [16] providing an overview of many of the existing approaches as well as a model to understand their relevance when applied to the customer churn task. Churn retention itself is further analysed in A.4 as a specific instance of survival analysis and time-to-event prediction.

A.3. Survival Analysis statistical methods

Among the most popular statistical models employed in survival analysis we can mention:

A.3.1. Kaplan-Meyer Approach

One of the most frequently used methods of survival analysis is the Kaplan-Mayer approach (KM). The KM method estimates the likelihood of survival [57]. One of the most relevant assumptions is that censored datapoints would have similar outcomes to non censored ones, and those recruited later in the study have the same probabilities as those recruited earlier. The KM plots a survival curve, which often reports median survival time(s), a reliable estimate if the majority of the observations are uncensored. One can calculate confidence intervals (CIs) for KM probabilities and plot CIs in the survival curves to provide a range of possible values for the population based on the sample. The curve itself does not provide information on whether or not the difference between the groups is significant. To do this, one can use the log-rank test [31].

A.3.2. Log-Rank

The log-rank test is a non-parametric test that compares two or more groups' survival distributions [87]. The aim of this test is to examine the null hypothesis: that there are no differences in distribution between the analysed groups. In such an application we can for example make an assessment whether the event rate in one group is higher or not compared to the other one.
The log-rank test is utilised when we are dealing with censored observations, while in their absence we can instead use the Wilcoxon rank-sum test to compare survival times [100]. The log-rank test itself is limited in the sense that it cannot determine an estimate of the difference between groups, whereas a Cox proportional hazard model can [83].

A.3.3. Cox Proportional Hazard Method

The Cox proportional hazards model is a semiparametric regression model that allows researchers to examine the effects of multiple variables on survival curves [100]. Semi-parametric means that the method does not require a specific distribution of the survival function; however, it does assume a relationship between the covariates and outcome [83]. The output of the Cox model is presented in hazard ratios (HR).
The term 'hazard' refers to the risk of an event occurring per unit of time, conditional on the subject having survived up to that time. The model is formulated in terms of the hazard function, $h_i(t)$, as follows [100][87]:

$$\log h_i(t) = \log h_0(t) + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni} \tag{A.1}$$

where $h_0(t)$ is the baseline hazard, while x_{1i}, x_{2i} ... are the measured baseline values for subject i associated with β coefficients, called hazard ratios. If an HR is significantly >1, then we can conclude that an increase in that risk factor corresponds to an increase in the event hazard, which in turn decreases the length of survival [87]. Cox proportional hazards model investigates relationships of predictors in these analyses and develops HRs. An important assumption made by the Cox model is that the proportional hazards between the variables remain steady over time. Cox hazard models allow researchers to adjust for cofounders and to form relative risk (RR) for individuals to experience an event based on risk factors.
From the general formulation of survival analysis we can further define the problem in some business cases, like for example in cases of churn prediction.

A.4. Churn prediction

The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers stop doing business with an entity. It is most commonly expressed as the percentage of service subscribers who discontinue their subscriptions within a given time period. In the context of survival analysis, churn can be seen as the event at the centre of the analysis and can be identified by the end of a contract or subscription. Compared to regular survival analysis the main difference is that the time t is set to a fixed value and the goal is to analyse if the customer will continue doing business (i.e. it will stay in the data set) after time t. Such a formulation consists of a binary classification problem of whether a customer will churn or not. Finer-grained analysis can be performed on the expected churn time, that is the length of time after which the customer will discontinue their relationship with the company, which is framed more as a regression problem.

A.5. Use case: customer churn retention in fashion e-commerce

Shopping, as many other aspects of everyday's life, is becoming faster and more instinctual as customers start demanding the chance to shop anywhere and anytime [59], with the comfort of buying products and having them delivered seamlessly in their homes. This change in the dynamics of purchases has deeply affected the structures and approaches in the retail sectors, inherently giving way to the diffusion and later on the widespread penetration of E-commerce. E-commerce itself involves buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, predominantly the Internet [13]. In its inception E-commerce lays out enormous possibilities of growth for the market [5], in particular for Small and Medium Businesses (SMB), which nowadays provide more than 50% of all of Amazon's products [7]. On a similar note we cannot avoid to mention that SMBs and in a more specific scope Small and Medium Enterprises (SME) represent about 90% of businesses and more than 50% of employment worldwide [10], which just means that the potential of growth that e-commerce could be able to provide is truly relevant and worthy of attention.

The adoption of e-commerce solutions can significantly improve the possibilities of growth for any company, especially those operating in international settings [89], and this, in conjunction with the general increase in online purchases and spending caused by the Covid pandemic [91], has given a decisive boost to the ever increasing shift towards online shopping. Thanks to this, research poises the overall growth of online purchases to reach

95% of the total by 2040 [37], while for now 2021 retail E-commerce sales amounted to approximately 4.9 trillion U.S. dollars worldwide. This figure is forecast to grow by 50 percent over the next four years, reaching about 7.4 trillion dollars by 2025 worldwide [21].
The prospect in such a market can only be promising for new business ventures trying to make their success, especially the ones operating in the fashion sector. Fashion is the largest B2C E-commerce market segment and its global size is estimated at US$752.5 billion in 2020. The market is expected to grow further at 9.1% per year and reach a total market size of US$1.164,7 billion by the end of 2025 [80].

In a deeper understanding of the infrastructure allowing for E-commerce to flourish it is relevant to understand which are the main advantages and differences compared to more traditional practices [86][32]: buying 24/7 all year long, convenience and time saving, possibility of comparing prices, availability of all information and details as well as easy navigation between categories and products. For companies this also translates in relevant advantages: no boundaries on the geographical scope of sales, lower operational costs and better services, general better returns on ad utilisation and the possibility to dynamically change policies and address distinct market needs.
On the other hand there are also some disadvantages to be mentioned, mainly: inability to try out items before buying them, the need for a basic understanding of electronic devices through which to conduct the purchase and finally problems related to delivery and shipping.
Given that the market is still far from being saturated, with 46% of US small businesses not having a website despite the average yearly growth of the E-commerce industry by 23% [47], many companies are still going to benefit from enacting steps towards a more online-oriented sales policy. In this environment it proves relevant to make use of all available advantages, chief among them the amount of data about customers, with customer experience and personalisation being the drivers for better sales and performances [14].
All of those factors have to either be carefully assessed or exploited in order to achieve optimal efficiency, but some of them represent significantly more acute issues for fashion compared to other categories of products. Fashion retail E-commerces generally suffer from higher than average shopping cart abandonment rates, around 40% of the total [82]. This is mainly caused by a lack of transparency on costs on the side of companies, that undermines trust in customers [48]. At the same time, the growing competition in the E-commerce environment is making customer acquisition processes and customer retention processes much more expensive than before, with industry stalwarts seeing their CAC up 70 to 75% whereas new markets are seeing increases closer to 50% over the past five

years [74]. In order to mediate such an increase in costs, that in some cases surpasses the average increase in revenue that the same companies realise, one of the most relevant solutions is to focus and invest more in customer retention processes. It's shown that it costs five times as much to attract a new customer than to keep an existing one [49] and that a customer who makes multiple purchases spends on average four times as much money as a customer who makes only one purchase [3]. Following this trend it is then relevant to start considering the CLV [25] as a measure for customer retention and a good metric for it.

A.6. Ethical challenges in fashion e-commerce

There needs to be an ulterior distinction within the fashion industry itself: that is caused by the gradual rift creating between what is today referred to as fast fashion, and its counterpart, slow fashion: according to [79], for every five new garments produced each year, three garments are disposed of. Incredibly, research has shown that 90% of our clothing is thrown away before it needs to be [26]. This is due to the increasing trend among fashion market leaders to sell as much as possible with as low a price as possible, i.e. fast fashion. This translates to enormous amounts of pollution and costs both from the environmental perspective (with entire rivers being rendered toxic by the chemicals used during production) and the working one (with most of the clothing being produced by workers whose wages are below the poverty line [19]). With the textile market on track to account for 25% of the global carbon budget by 2050, retailers and consumers alike are experiencing an awakening [18].

In an effort to increase the sustainability of such an industry, there has been a gradual emergence of an opposite trend to this one, aiming at making the whole industry more sustainable by reducing overall volumes of sales and making products that would last longer, along many other initiatives. Such developments fall under various denominations, one of them being ethical fashion: ethical fashion is the designing and manufacturing of clothes while caring for the people and communities involved in the process, and while also minimising the impact on the environment. It focuses on both the social and environmental impact of fashion, seeking to improve the working conditions of labourers and the environment [64]. Such a concept gets even further expanded under the denomination of slow fashion, that instead is an approach to producing clothing which takes into consideration all aspects of the supply chain and in doing so, aims to respect people, the environment, and animals. It also means spending more time on the design process, ensuring that each piece of apparel is quality made [65].

In the end the environmental and business issues boil down to a category of fashion companies aiming to reduce overall consumption and growth [78]. In order to ensure profitability and maintain competitiveness under such premises the solution is trying to provide the highest quality of service possible and in doing so ensuring a better customer retention. Given the relevance of customer experience, it then proves worthwhile and extremely useful to include some data science solutions to the problems, that may help in better categorising customers and understanding their behaviour. This links perfectly with churn retention improvement, given that figures show that a decrease of just 5% is enough to drive an increase in profits of 25-95% [34]. Personalization is key in such settings, given that sending out non targeted initiatives may instead risk doing more harm than good. It follows that a targeted strategy is to be preferred [38].

B | Interpretation framework for the churn prediction task

This section introduces a framework that will help connect the results of the survival analysis task to the actual meaning they have in the analysed context. Such a framework is developed leveraging on business expertise and support from ASKET itself and will now briefly be introduced and explained.

The aim of obtaining an accurate churn prediction is to get with reasonable certainty an indication of whether an event will happen and in our specific case of TTE prediction, when that is estimated to happen.
From the interpretation of churn given in section A.4, such a task is often analysed as a classification problem rather than a regression one. This inherent lack of immediate correspondence sparks the need to clarify how a TTE prediction can effectively be interpreted in a churn framework. The possible prediction results in the surface can be two: either a customer is predicted to churn or not. Establishing the possibility of not churning foresees the chance the customer will interact with the company again in the future and as such, this event can be estimated through a regression task.
Another very important factor that emerged throughout the analysis is the time at which the actual prediction is performed. It can happen in fact that through the TTE prediction (defined as the amount of time transpiring between the last event and the next one, not between the time of prediction and the next one) ends up indicating a date that is preceding that of the algorithm execution. Such a case in itself is not usually contemplated in churn analysis, since it would inherently be classified as a customer having *already churned*.

Based on the specifics of churn in such an application case we can thus try and formalise it by describing an immediate and easy to understand framework. Such framework will consider two main dimensions to operate:

- *Time Of Event (TOE)* indicating the date when the regression task predicts the next event to occur

- *Execution Time (ET)*, indicating the date when the prediction algorithm itself is executed

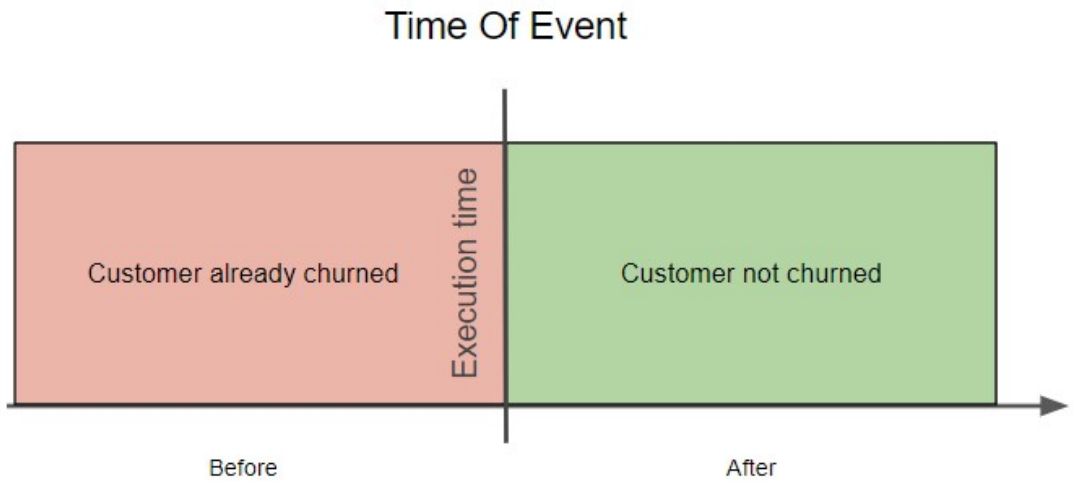


Figure B.1: Churn interpretation framework.

As we can observe in Figure B.1, based on the two dimensions at our disposal, TOE and ET, we can safely assess whether a customer is predicted churned or not at an initial glance of the results.
A prediction of TOE on the left hand side of the graph, that is before the ET ($TOE \leq ET$) means that according to the model, the event (purchase) should have happened before the date of execution. Given that this has not happened, based on the model's prediction we can assume the customer to be *churned*. On the other hand whenever the TOE is predicted after ET ($TOE > ET$), there is no reason the customer is already churned given that the purchase event still has to happen both in actuality and in prediction.

Since churn retention has been traditionally been addressed differently under the guise of a classification task (as presented in section 1), framing it in a TTE framework requires to lay down some groundwork to better understand how to interpret results. The actual framework ends up being rather simple and immediate: it only leverages on two characteristics, the TOE and ET and from that allows to establish how from a TTE prediction it is possible to assess whether a customer has actually churned or not. The main assumption to be made is the certainty of an event happening in the future for each data point (customer order history) in the analysis: while this could be true in instances where churn retention is applied to subscription or contract based services, with the event being the cancellation of such a service, whenever online purchases are the subject this is not as easily applied, as a customer does not formalise the end of his purchase interaction with a company. Such a problem resides in the fact that the lack of an event is difficult

to encode, given that no specific amount of time could be hypothetically too large to safely say that a customer has churned or not. Such a step can only be formalized in the interpretation of the results.
In order to accommodate to the last cited limitation the proposed and adopted solution in this thesis is to proceed in the TTE analysis without including customers belonging to the "Churned" cluster. In this case there is reasonable certainty that such customers will not place a purchase in the near future and thus including them in the analysis only represents them as a challenge to the interpretation framework.

The exclusion from the analysis at a given point though is not final, since the placing of a purchase after the analysis is conducted may instead see them shift their affiliation to another cluster once the clusterization task is run again.
On the other hand practical application has shown that such a limitation has to be very rarely enforced, since no member of the "Churned" cluster satisfies the demand of having placed at least 3 purchases in the analysis window.

C | Data Mining extended

After an initial assessment of the available data and the computation of the highest possible amount of features for the analysis, the next step has been to try and reduce the overall amount of features both to achieve a less complex data representation and more consistent layout.
The first step in the analysis has been in considering the use of some dimensionality reduction techniques, like PCA. The benefits provided by such techniques mainly reside in the capability of assessing the explicative power of attributes in a dataset by evaluating their "explainable variance", that is how much of the variance of the data can be provided by each single attribute. Further and more detailed references are provided in 2.3.1. Given the mixed nature of the data available for the analysis, both categorical and numerical, a technique like PCA cannot be utilized. At the same time an alternative technique called FAMD allows to perform the same kind of analysis on a set of attributes of mixed nature. This techniques is more deeply analysed in section 2.3.2. The explainable variance of the various attributes has been calculated on various possible values of the number of components to be utilised and has yielded results as shown in Table C.1.

Table C.1: Explainable variance provided by varying number of principal components obtained from FAMD analysis.

Number of Components	% of Explainable Variance obtained
2	16.53
10	46.11
20	67.03
30	83.58
40	96.28

As can be observed by the results of Table C.1, the explainable variance provided by the components identified through FAMD is definitely not enough to perform any meaningful data dimensionality reduction. We can observe that an equivalent of 30 components would be needed to go over 80% of explainable variance of the original dataset. Such results lead to believe that a dimensionality reduction task following this approach is not worthwhile

and that other solutions should be utilised.
Such a result can be further be appreciated when looking through the distribution that datapoints assume when plotted in a space that puts them in relation to the Principal Components identified through FAMD analysis.

Given the diverse nature of the available attributes and the high diversification that they are able to provide to the dataset, it proves useful to get a deeper understanding of the most relevant ones and perform an iterative process of exclusion through which lesser attributes can be isolated and removed from the analysis. An initial starting point for such considerations is in the overview provided to us by the FAMD analysis on each feature's contribution to the dataset's variance, showcased in Figure C.1.

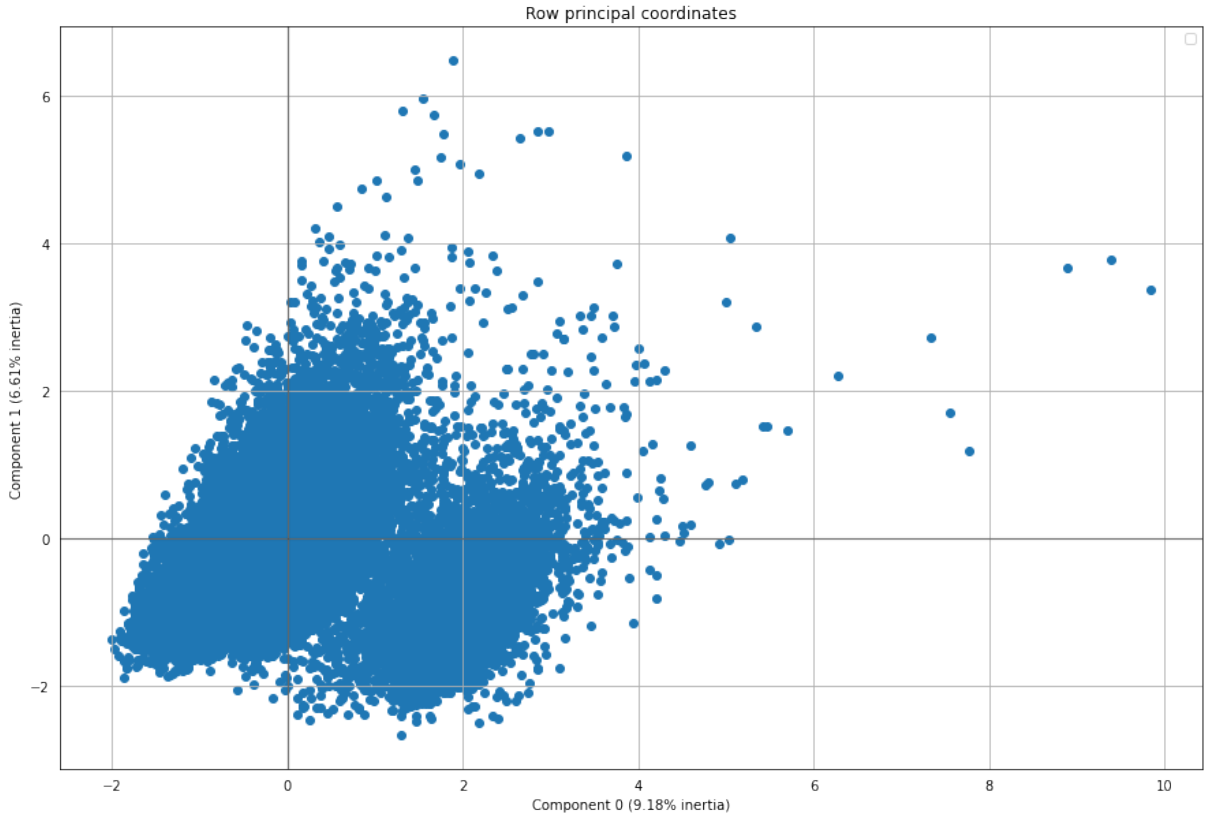
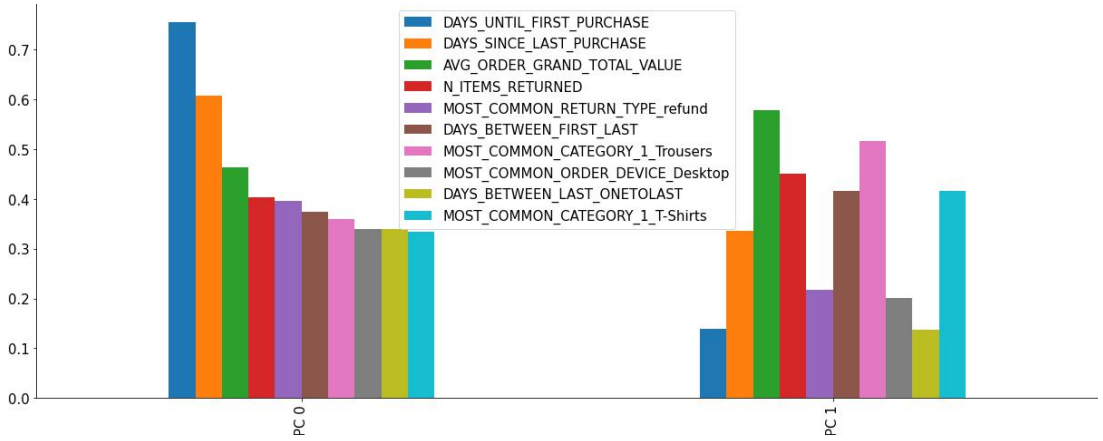


Figure C.1: Plot of dataset along the first two FAMD principal components.

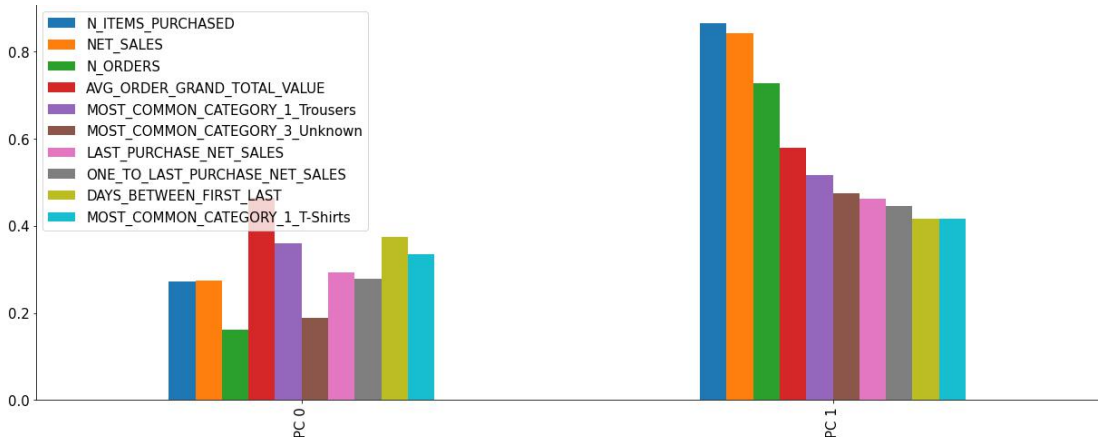
In Figure C.1 we can find the plot of all datapoints according to their coordinates along the first and second principal components observed through the FAMD analysis. As we can observe there is a slight hint of separation between those that could look like two major clusters in the dataset as well as many different outliers. The lack of a clear distinction and the lack of promising results in the ability of the dataset to be properly represented by a subset of features as defined through the FAMD analysis is unfortunately an additional reason that can lead us to disregard the principal components and proceed with further

analyses.

In order to obtain some insights from such an analysis and try and link it with other kind of information to try and optimise the feature set, we can start off by checking how features contribute to the composition of each principal component in Figure C.2.



(a) Features ordered by explained variance contribution to the first Principal Component.



(b) Features ordered by explained variance contribution to the second Principal Component.

Figure C.2: Features explained variance contribution to Principal Components 0 and 1.

For convenience sake only the 10 most relevant feature have been plotted according to the contribution provided to the first component in C.2a and to the second one in C.2b. On the left hand side we can observe the plot of contributions to the first component, while on the right side we observe the contributions to the second component of those same features. From the overall analysis of the results we can observe that the first component (able to encompass 9.18% of the dataset explainable variance) is heavily made up of features related to the time intervals between orders (represented by features like "DAYS_UNTIL_FIRST_PURCHASE"), return order information, category of products

purchased and finally devices through which the order has been placed. The second principal component (providing 6.61% of the dataset explainable variance) is more mixed and showcases higher relevance for numerical order related features (Number of ITEMS PURCHASED, NET_SALES) and preferred category of purchased products. It is noteworthy to mention that web interaction features also managed to score high in the provided explainable variance, although suffering from being less populated than other features. This references back to Chapter 3.2.1 where the lack of such data for a big component of the customers is a criticality in the dataset.
All those considerations can be taken into account when trying to identify a less formal and more contextualized assessment of the variables that could lead to exclude some of them because of redundancies and similar information. We can be aided in such a task by looking through the correlation matrix of the current dataset in Figure C.3 and from there starting to elaborate further.

In Figure C.3 we can observe that while on general terms many features show to be almost completely uncorrelated, we can identify some other more critical occurrences that could be added reasons for choosing one feature over another one. Let's try to go through the most relevant instances and then summarise the conclusions.

- *Number of Orders* vs *Number of Items Purchased*: given that the amount of orders is strongly correlated to the number of items purchased as results of such orders, it is useful to reduce the load in information giving preference to the *Number of Orders* feature because of its relevance to the ends of the adopted RFM framework
- *Total Sessions* vs *Sessions Last 6 Months*: given the recency of adoption of website interactions in ASKET's dataset it is very common for sessions in the last 6 months of analysis to overlap with all the ones ever recorded. Given the added generality of the *Total Sessions* parameter we are going to remove the less general one.
- *Total Sessions* vs *Sessions Containing ...* and *Products Clicked*: all of the latter features are further derivatives of the more general recording of the wider concept of session. This results in a high correlation between such features, although it may not prove substantially problematic given the different intrinsic nature of such events. Given the small availability of data that website interaction have for the overall dataset though, it does not make sense for this analysis to go in this much detail and we can safely exclude them.

Based on the insights gained from the correlation matrix and combining them with the observations from the FAMD analysis, a definitive list of parameters and features can be drawn, excluding those not proving to be informative enough to be included in the

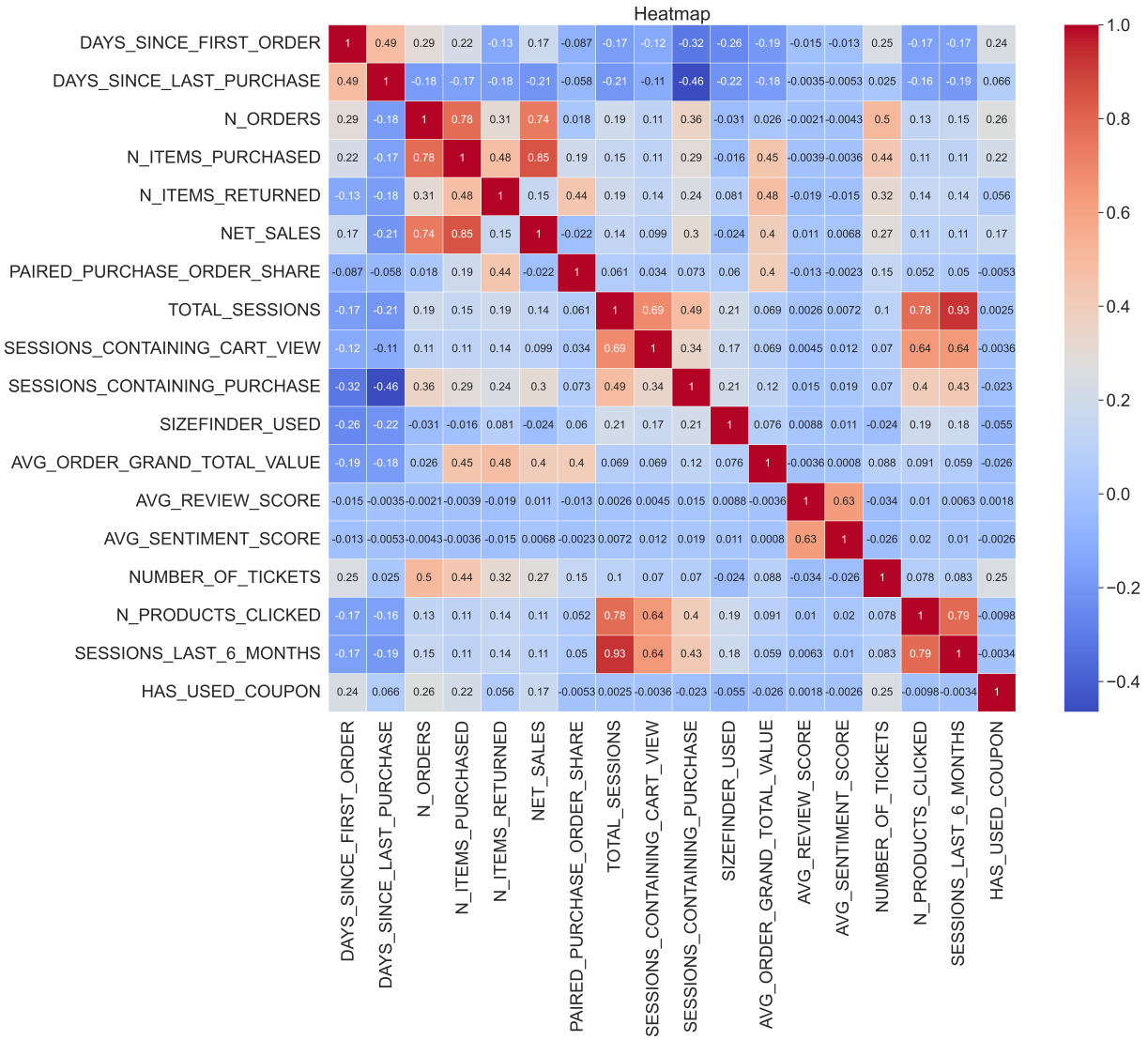


Figure C.3: Correlation matrix of all non-RFM related features.

current dataset without reasonably subtracting any relevant information to the model itself. To showcase the dataset Table C.2 will use the same 3-topic structure introduced in the beginning of section 3.2.1.

As can be observed in Table C.2 aside from the features already listed for removal in the correlation matrix analysis, all the other removed ones can be traced back as being some of the least relevant in the FAMD analysis and being redundant with the final ones. It is worth to mention that none of the RFM variables have been removed according to the decision to maintain such a framework into the analysis.

We report briefly the removed attributes for reference: Most Common order Device, Average cart views, Item level Return rate, Days since First order, Most Common Category 3, Most Common Payment Method, Paired Purchase Order Share, First Purchase Source,

Table C.2: Final data configuration.

Orders and Returns	Order Interactions	Web Interactions
Market	Most Common Return Type	Total Sessions
Exchanges per Order	Most Common Category 1	Pageviews Per Session
Days from Last Purchase	Most Common Category 2	Most Common order Device
Number of Orders	Sizefinder Used	
Number of Items Returned	Coupon Used	
Net Sales	Number of Tickets	
Most Common Delivery Method	Predicted Gender	
Average Order Value	Most Common Collection	

Average Review Score, Average Sentiment Score, Share of Other Products.

Now that the final analysis dataset has been consistently defined we can also address the problem of missing values. As we have mentioned there are features that present a lack of data because of their collection starting after that of the initial attributes. Aside from this there are occasional orders with missing or corrupted information that has to be conciliated in order to make the analysis possible. Such a process was already introduced before conducting the FAMD analysis, but for the sake of conciseness we are going to discuss about it now referencing directly those features that have not been removed from the attribute set.
In most cases regarding numerical variables the decision taken has been to influence the dataset with as small an artificial variation as possible, by substituting the same value with the mean average of that attribute calculated over the non null values. In the instance of categorical variables, where the criticalities only concern attributes like *Most Common Delivery method*, *Most common Return type*, *Most Common Category 1-2*, we adopt the following type of conciliation:

- *Most Common Delivery Method*'s not clearly defined or corrupt values will be changed to the category "Other"
- *Most Common Return Type* presents many exchanges or returns without a specific indication of the kind of solution adopted by the client, thus they are hard set to "Unknown" already in the dataset
- *Most common Category 2* is the only attribute out of the item category ones that could have a nulll value in the instance for example of customers having bought the same category of items in all of their orders. In these situation it gets set to "Unknown"

D | Preliminary Data Analysis

In this section we provide an overview of data as it is available from the source. The most relevant attributes will be addressed and commented and indications of data distributions with some preliminary considerations on the clusterization task will be provided.

Given the presence of both categorical and numerical attributes in the dataset, we will address their description separately. In the case of numerical attributes it is worthwhile to analyse the interval of values, as to make the future understanding of clusters easier when comparing their distribution compared to the general behaviour. Such data will be summarised in table D.1.

Table D.1: Preliminary assessment of numerical attributes distribution and interval of values.

Feature	Mean	Max	Minimum
Days From last Purchase	471	1	2427
Number of Orders	2,16	1	62
Number of Items Returned	1,1	0	155
Net Sales (SEK)	2817	0	144184
Average Order Value (SEK)	1706	120	43500
Exchanges per Order	0,01	0	9
Number of Tickets	0,47	0	25
Total Sessions	3,96	0	320
Pageviews per Sessions	5,5	0	160

Just as a quick reference to the data provided in table D.1, we can observe that the *Days From Last Purchase* mean value is very high, considering that we are accounting for the whole dataset, including already churned customers that have not placed purchase in the last few years after their first one. The same explanation can be given for the high max value.

A particular mention has also to be done with respect to *Net Sales*: such an attribute is expressed in Swedish Krona (SEK) and it is computed, as referenced in table 3.2, as the value of sales minus that of exchanges, meaning that customers may potentially return

all items and net a zero profit for ASKET.
A final clarification is to be provided for the Web Interaction attributes *Total Sessions* and *Pageview per Session*, for which the mean is computed only on customers that have a non-zero value. As already mentioned, only a portion of customers are tracked in their online interactions and as such it is more meaningful to provide the average of the informative values.

With respect to the categorical features making up the other set of attributes in the dataset a very brief view of their distribution and most common values follows in Table D.2.

Table D.2: Preliminary assessment of categorical attributes distribution and most common values.

Feature	Most common value	2nd most common value
Market	Sweden ($\sim$ 25%)	Germany ($\sim$ 21%)
Product Categories	T-Shirt ($\sim$ 35%)	Trousers ($\sim$ 34%)
Product Collections	Man ($\sim$ 84%)	Woman ($\sim$ 13%)
Order devices	Desktop ($\sim$ 19%)	Mobile ($\sim$ 16%)
Return Types	Unknown ($\sim$ 68%)	Refund ($\sim$ 22%)
Delivery Methods	Regular ($\sim$ 76%)	Express ($\sim$ 18%)
Coupon Used	No ($\sim$ 97%)	Yes ($\sim$ 3%)
Sizefinder used	Yes ($\sim$ 84%)	No ($\sim$ 16%)
Predicted gender	Male ($\sim$ 80%)	Female ($\sim$ 18%)

As per the categorical features in Table D.2, it is possible to start drawing an initial understanding from the dataset. Being that the only working assumption is that the more an attribute is common, the more its relevance to a customer actually conducting a purchase is high, it is still interesting to analyse the composition of the customer base for the future purposes of the clusterization task. At the same time, the analysis of single attributes is not going to provide us with that much insight, thus there will be a listing of the most prominent correlations between said attributes in order to draw a more comprehensive image of the overall dataset.
Most customers hail from Sweden and Germany, as Sweden is the longest served country by ASKET and most of its customers are the oldest ones in the dataset. It is also interesting to point out that one of the fastest growing customer bases is in the US, with extra-European customers obviously having the potential of behaving differently than more consolidated and older European ones. Finally, Germany has become as of now the second most relevant European country served by ASKET, with German customers representing a more mature and higher spending customer base.

At the same time it is also curious to observe that while the T-shirt remains the oldest and most famous garment by ASKET, it was mostly bought in the initial years of the company, while now newer customers tend to buy more expensive apparel, like the trousers.
While men have been the traditional target of ASKET's garment collection, with women apparel making its appearance in the store only in May 2021, most customers have thus been men, with women starting to become more relevant only recently and up until now maybe interacting with ASKET for gifts, rather than purchases for themselves.
It is generally more common, at least among tracked customers, to place purchases through a desktop device, which provides a better overview of the website, of garments and generally offering a better customer experience.

List of Figures

List of Tables

Acknowledgements

I would like to thank my academic supervisor, Marcello Restelli, for the guidance and support offered during the thesis development.

I would also like to offer my deepest thanks to my industrial supervisor, Vidar Trojenborg, and the whole of the host company's, ASKET, team for the wonderful opportunity of working alongside them as well as the familiar and welcoming atmosphere characterising such experience.

I would then like to offer my thanks to Federico Schiepatti, EIT reference for PoliMi, who has been a constant point of reference during the length of my double degree experience and without whom everything would not have been as smooth and achievable.

I would now like to dedicate some special acknowledgements to family and friends that have been of instrumental help all throughout the thesis development and in general along the duration of my wonderful university years. It is also thanks to them that I have reached such a huge milestone and I am so glad to have been able to count on their unwavering and continued support. My deepest thanks go:

- To my wonderful mother Angela, gone too soon back into the arms of our Lord, who through her powerful presence as an emblem of dedication, passion for one's work, perseverance and kindness towards everyone has helped shape me into the person I am today. This thesis is only one of the many steps that I hope to accomplish as a testament of the love and affection she poured towards me as a mother;
- To my dear father Massimo, that did not manage to attend either of my graduations as he joined his wife after so much sorrow. He has been the pillar on which to lean on whenever things were rough and remains a yearned for memory in every instant;
- To my sweet grandmother Anna, who played the roles of mother and father for me for so many years and has been the enabler of my university experience, my travels and my serenity. As only she could have known the joy of, this accomplishment is

also hers, for she is the reason I managed to endure through the many sorrows that life has put in my way until she also went away too soon;

- To my grandfather Giuseppe, that despite the many years we have not been sharing together, still remains a role model of hard work and perseverance against any obstacle for me. With the thought of how wonderful celebrating with him would have been, I thank him from my heart;
- To my grandmother Marisa, who despite suffering the distance to which studying far from home has subjected us, always supports me, cares for me and wishes the best for me. I thank her for her affection and how she makes every moment back home sweeter and happier;
- To my aunt Ornella, that has always supported me in my study path and has always managed to make me feel the familiar warmth and closeness of staying among dear ones despite the distance;
- To my uncle Giuseppe, that has always encouraged me to pursue and go further in my studies, always interested and keen to know what my current interests were, where the hot topics lied and what would be next;
- To my cousin Aurora, that with her affection and simple vicinity has been for me at times a close friend, a dear relative, an understanding supporter and most of all, more than a sister in times of need;
- To my cousin Viola, whose admiration over the years has encouraged me to always push further and aim higher, with the hope that by now she sees her full potential and can more than surpass me in her own university path towards ever higher goals;
- To my uncle Fabio, that has been for me very close and affectionate in many difficult moments and that despite everything has always had a deep care and thought for me;
- To my dearest friend Oscar and all of his family, that for many years now has been for me a source of advice, a great confidant, a honest friend and the one person who I could always rely on, no matter the circumstances. His affection for me has only been amplified by the care of all of his immediate family, that I always keep in my thoughts and thank for the wonderful moments spent together;
- To my closest colleague Andrea, that throughout our university journey has grown to be one of my dearest friends and with which I have had the pleasure to work on so many occasions. I thank him for his knowledge, his judgement and his unconditional

support in every step of my Master's path;

- To my dear colleague Stefano and the whole of Alpha Team, along Leonardo and Marco, that have made my years in Politecnico unique and such a better learning and growing experience. With the hope of crossing our paths soon, I wish to thank them for all the time spent together;
- To my companion of any university years Mattia, with which I have shared a roof for the whole length of my years in Milan and that has had the patience to stand me and the sharp mind that has helped me raise so many questions about my professional and academical development. Wishing him the best in his own studies, I thank him for his constant support;
- To my dear friend Federica, that I thank after being friends for all of these years for the support she has always been ready to provide. Along being a close friend, she has always helped me discuss and think about so many crucial developments that have seen me move from a teen straight out of high school into a more realized and complete person that by now has reached the end of his academic adventure. I thank for having been there, with the hope of celebrating as soon as possible her own graduation.
- To my friends Sara and Giulia, that throughout the years have remained faithful friends and a certainty in my life. Thanks are due to them for having been so important in supporting me throughout many hardships and some of the few people I always hold dear to my heart wherever I am in the world.
- To all of the new amazing people I have got to know in Stockholm and with which we have managed to build a small, but stable friend group that has been instrumental in making my stay far from home more enjoyable and joyful. It is through their support that I have made it through my thesis process and all of the hardships it included. Thanks for this with the hope that the future will still hold many great moments for us to share.

Thanks to everyone that I have not had time to mention here, but with which I have shared struggles, joys and important moments together. You have all been so important for me.