POLITECNICO DI MILANO

School of Industrial and Information Engineering
Management Engineering

AUTOMATED MACHINE LEARNING: COMPETENCE DEVELOPMENT, MARKET ANALYSIS AND TOOLS EVALUATION ON A BUSINESS CASE

Supervisor: Prof. Barbara Pernici

Correlator: Alessandro Volpe

Master Degree by:

Tommaso Curotti 905582

Academic Year 2018/2019

Abstract

Today data are increasing in volume and in complexity every day, to follow this exponential growth are necessary new methods to analyze the enormous amount of data generated. Companies have understood that data are essential to compete, and they are collecting even more data, but sometime happens that they do not have the resources to analyze them. The actual problem is not how to generate data but how to analyze data. The real problem today for companies is the scarcity of expert data scientists able to extract value from Big Data. To solve this problem are arising new technologies of data management and analysis. One of the most potential innovation in this field is the automated machine learning. The thesis aims to define it considering different perspectives to increase knowledge and to test how these systems work.

Automated machine learning is the new technology enabling the development of machine learning models autonomously by the machine. These systems reduce the level of interaction human-machine, letting the machine to decide how to create complex models to solve complex problems. Machine learning is in the hand of few skilled people, data scientists, having deep knowledge in computer science, mathematics and statistics. With AutoML systems this equilibrium is wrecked since even persons without strong skills can rapidly develop machine learning solutions.

To create a comprehensive ‘big picture’ of this world, different analysis both qualitative and quantitative were executed. Qualitative researches were carried about understanding what automated machine learning is, researching the reasons why this new technology is becoming even more requested by companies and studying the potential impact that it could bring to data science professional figures in doing their jobs. Quantitative analysis was about both quantification and classification of the actual AutoML systems, trying to define the number of different solutions in existence today and their target customers.

Always regarding the quantitative researches were tested different automated machine learning systems, precisely five, comparing their performances with those one of two different machine learning solutions really implemented to solve a prediction problem for a gas dispatching company. The comparison was conducted considering a specific trade-off composed by three main drivers: MAPE as metric to establish the goodness of the different models compared; the economic return, directly correlated to MAPE, brought by the different solutions in order to understand if AutoML can achieve the same results obtained by traditional machine learning; the set of time, cost and effort needed to develop the different model of prediction both for AutoML and traditional machine learning. This final part of the thesis aims to answer to the question: could AutoML be considered as a valid alternative to traditional machine learning workflow? As will be seen from the conclusions drawn from the tests carried out on the business case under consideration, this new technology drastically reduces development time and costs, but does not always respect the level of performance desired. The choice whether to adopt these systems or not is based on the triple trade-off performance-time-development costs.

Abstract (italiano)

Ogni giorno i dati stanno aumentando di volume e complessità, per seguire questa crescita esponenziale sono necessari nuovi metodi per analizzare l'enorme quantità di dati generati. Le aziende hanno capito l'importanza dei dati per competere nel mercato e per questo stanno aumentando la quantità di dati raccolti, ma capita che sempre più spesso non dispongono delle risorse per analizzarli. In effetti, oggi il problema non è come generare dati ma come analizzare i dati. Il problema per le aziende oggi è la scarsità di esperti data scientist capaci di estrarre valore dai Big Data. Per risolvere questo problema stanno nascendo nuove tecnologie di gestione e analisi dei dati. Una delle innovazioni col più grande potenziale in questo campo è il machine learning automatizzato. La tesi mira a definire questa tecnologia considerando diverse prospettive per aumentare la conoscenza su di essa e testare come effettivamente questi sistemi performano.

Il machine learning automatizzato è una nuova tecnologia che permette alla macchina di sviluppare autonomamente modelli di machine learning. Questi nuovi sistemi riducono il livello di interazione uomo-macchina, permettendo alla macchina di decidere come creare un modello di machine learning per risolvere problemi complessi. Il machine learning è nelle mani di poche persone, i data scientist, i quali hanno profonde conoscenze in computer science, matematica e statistica. Con i sistemi di AutoML questo equilibrio potrebbe vacillare, dato che anche persone senza forti competenze possono rapidamente sviluppare soluzioni di machine learning.

Per creare una comprensiva 'big picture' di questa tecnologia sono state eseguite diverse analisi sia qualitative che quantitative. Le ricerche qualitative sono focalizzate nel definire il machine learning automatizzato, ricercando le ragioni del perché questa nuova tecnologia stia diventando sempre più richiesta dalle aziende. Inoltre, è stato oggetto di ricerca indagare sul potenziale impatto che l'AutoML potrebbe portare nel metodo di lavoro dei data scientist durante il tipico processo di sviluppo di un modello di machine learning. Le analisi quantitative, invece, sono state condotte al fine di classificare e quantificare i sistemi AutoML presenti oggi nel mercato e per definire quale sia la loro strategia di targeting.

Sempre riguardo le ricerche quantitative sono stati testati diversi sistemi di machine learning automatizzato, precisamente cinque, comparando le loro performance con quelle di due soluzioni di machine learning tradizionale realmente implementati per risolvere un problema di previsione per una compagnia che si occupa di dispacciamento di gas. La comparazione è stata fatta considerando uno specifico trade-off composto da tre driver principali: il MAPE come metrica per definire la bontà dei diversi modelli confrontati; il ritorno economico, direttamente correlato al MAPE, portato dalle diverse soluzioni per capire se l'AutoML può ottenere gli stessi risultati ottenuti dal machine learning tradizionale; l'insieme di tempo, costi e sforzo necessario per sviluppare i diversi modelli di previsione sia delle soluzioni automatiche che di quelle tradizionali. Questa parte finale della tesi cerca di rispondere alla domanda: l'AutoML potrebbe essere una valida alternativa al tradizionale metodo di lavoro per sviluppare modelli di machine learning? Come si evincerà dalle conclusioni tratte dai test effettuati sul business case preso in considerazione, questa nuova tecnologia riduce drasticamente il tempo e i costi di sviluppo, ma non sempre rispetta il livello di performance voluto. La scelta se adottare questi sistemi o meno è basata sul triplice trade-off performance-tempo-costi di sviluppo.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

In recent years the world has seen an exponential growth of machine learning (ML) techniques in many fields and industries. However, ML methods are very sensitive to a wide range of variables and design decisions, which are a considerable barrier to new users of these techniques. ML can be split into many subdivisions covering a very large territory of knowledge and competencies, which need deep knowledge in many fields. To simplify the adoption of ML was born the automated machine learning (AutoML) that aims to automate the entire pipeline from data preprocessing to model selection.

AutoML is the core topic of this thesis. It is a very new hot topic in the field of Artificial Intelligence (AI) that could bring massive changes in the various methods companies and professionals work with data and develop high-value ML models. The idea behind AutoML is to automate ML pipelines, where pipelines are all the sequence of steps that a traditional project of ML has to pass through to achieve its results. Ideally, AutoML is an extremely powerful tool that enables everyone to conduct analysis and project on data even without knowledge in programming, mathematics, and statistics. AutoML can lead to improved performance of the models while reducing the time and effort spent by data scientists. As consequence, AutoML gained a real commercial value in recent years and several big companies are developing their AutoML systems to sell in the market. The purpose of democratizing ML is not to impute only to proprietary tools but also to open source tools, both deep analyzed in the chapter of the market analysis.

In the following chapters we will go through different aspect of AutoML, trying to describe how it works and what it is changing in the field of ML, then the impact that it has on the market, organization, professional role and to conclude we will analyze all these aspects in a very insightful business case developed in Bip. xTech. The business case will show the differences between a predictive ML project for an oil & gas company and then it will compare the same project developed with five different proprietary AutoML tools tested in Bip, to build a useful benchmark and point out the benefits and the constraints that AutoML has over ML. The tools used to develop the business case are Google Cloud AutoML Tables, Dataiku, Azure ML Microsoft, AWS Sagemaker and H2O Driverless AI.

1.1 Objectives

Being AutoML very new there are few types of research about it and a difficult part of this thesis was to define its objectives. The difficulty was given by the type of analysis to conduct: the indecision was about to develop a technical thesis covering the technicalities behind the implementation and construction of an AutoML system or to highlight AutoML impact in the market, in the organization, on the performance, and professional data science roles. Thus, we decide to make an inside-out analysis that tries to describe and build an overview of what an AutoML system can do and then go to analyze the impact that it has on the external variables. So, to build an inside-out analysis we defined four principal objectives to achieve:

1. To create knowledge and awareness about a very new hot topic in the field of data science, automated ML. The research is not aimed to describe in detail the technicalities and algorithms behind the AutoML solutions, but to point out the main characteristics of these new tools and to explain which part of the traditional ML pipeline AutoML impacts.
2. To understand how AutoML is changing the world of data science, in this perspective the focus will go on the study of three main impacts:

    - Data scientist profession
    - Arising professional role: the citizen data scientist
    - Organization of data science workflow, considering the CRISP/DM model

3. To create the big picture of the actual tools both proprietary and open source based on some features selected to classify these tools. After having created the big picture of the actual market we will analyze the main customer base of the most important proprietary tools to understand which kind of industries are interested and are using AutoML solutions. This chapter will be concluded with a deep-dive on the tools used in the business case of the fourth chapter.
4. To show what are the differences between a traditional ML project and the same project develop with different proprietary AutoML tools. This part will be integrated and explained with a real case developed in Bip.xTech. The Business case will be structured as follow:

    - Description of the context in which the oil & gas company operates
    - First intervention of Bip and results
    - The second intervention of Bip and results
    - Description of the same project developed with AutoML tools
    - Comparison of the different solutions
    - Conclusions

The whole thesis is focused to achieve the first objective: to create knowledge and awareness about all what concerns this new field under different point of views. Instead for the other three objectives, there is a dedicated chapter that treats a specific argument. To achieve these four objectives the thesis will begin from the history of ML and why it is fundamental to deal with data. Then we will introduce the new branch of artificial

intelligence called AutoML. Introducing AutoML we will focus our attention on the birth of AutoML and then we will go in-depth about how it impacts the traditional ML pipeline, discovering what are the main functionalities, benefits, and constraints that AutoML brings with itself, firstly from a theoretical point of view and then from a practical perspective adopted in the Chapter 4 where we will analyze AutoML experiments.
Once we concluded the introduction part the focus will go towards the second section of the thesis that tries to satisfy the objective of understanding the impact that AutoML has on the professional role of data scientists and on the organization of the data science team. Since that, theoretically, AutoML can perform all the tasks that usually are executed by data scientists, it is interesting to discuss the possible future scenarios in data science and to describe the new role emerging thanks to AutoML. This part will be integrated with some important professional article analysis (Forbes, Forrester, and Gartner).

The third part of the thesis, treating the relative objective, is aimed to create a benchmark of the different tools. To build a concrete big picture of what kind of solutions we can find in the market today, we will discuss and analyze through a consistent market analysis all the available tools in the market nowadays. In order to comprehend that it is not just a theoretical topic but a real added-value for every company that deals with Big Data, during the market analysis, the major information that we will extract will be the typology of the different tools (open source or proprietary), who are the leaders of the market and all the technical functionalities of each tool analyzed. An insightful part of the market analysis aims even to map and classify the target customers for the main proprietary tools for which are available trustable information about their customers with relative business cases published on their websites.

The final part, aimed to satisfy the fourth objective of the thesis, is a practical comparison between traditional ML projects developed in Bip.xTech, a very successful case of study, and the same problem solved through the usage of different AutoML tools.
The different solutions will be analyzed under three main key performance indicators (KPI):

- MAPE (Mean Absolute Percentage Error)
- Economic return (potential incentive gained or lost)
- Project cost, time and effort

This comparison is very useful to comprehend in which measure AutoML can help companies and data scientists to do ML projects. All the details will be discussed in Chapter 4.

1.2 Big Data

Before starting to talk about automated ML is important understand why society, companies, people and everything is around us need to deal and comprehend a very important element: data. The main reason could be imputed to Internet. Internet is the reason of the technological evolution that lead humankind to develop new ways to deal with things around us, adopting new ways to communicate, to interact, to move and to behave.

So, in order to understand why we talk about automated ML we can start from Big Data that is one of the main consequences of Internet growth and enabling devices diffusion.

> *"Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time."* **[1]**

From the definition, Big Data are very huge amount of data that can't be analyzed with common tools because the time needed is too long and today actions must be taken faster. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big Data are created by every single person or device connected through Internet or able to collect and store data. Big Data are all those data on which is possible to conduct analysis trying to extract important information, they have a real value if managed efficiently. Usually, to define Big Data is used the 5 Vs model that explain the concept of Big Data with its 5 main characteristics:

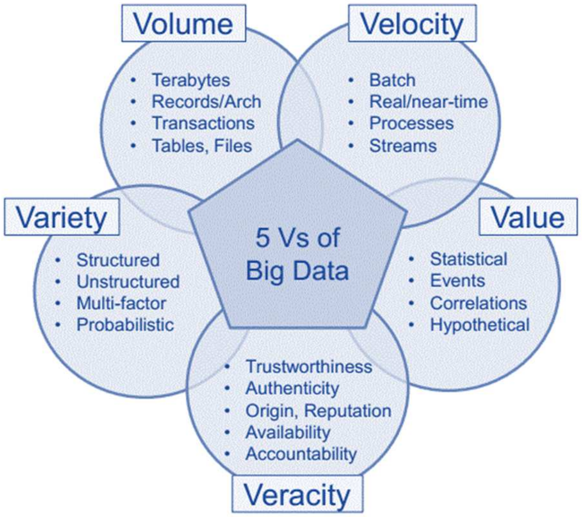


Figure 1 - 5 V of Big Data

Velocity: obviously, velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every day the number of emails, twitter messages, photos, video clips, etc. increases. Every second of every day data is increasing. Increasing the velocity of data generation means also increasing of the pace at which they are generated. This leads us to develop new method to handle this new dynamic situation.

Variety: variety is defined as the different types of data we can now use. Data today looks very different than data from the past. We no longer just have structured data (name, phone number, address, financials, etc) that fits nice and neatly into a data table. Today's data is unstructured. In fact, the Computer World magazine states that unstructured information might account for more than 70%-80% of all data in organizations. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

Variability: the meaning or interpretation of the same data may vary depending by the contest in which it is collected and analyzed. The value, thus, is not held by the data itself but it is strictly linked to the contest to which come from.

Volume: it refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc. The vast amounts of data have become so large in fact that we can no longer store and analyze data using traditional database technology.

Value: it refers to the real value of the Big Data, they record significant events with several types of data correlated from each other. From the analysis of these data is possible understand why events happens and which are the important factors related to events.

Veracity: Veracity is the quality or trustworthiness of the data. It is the degree to which data is accurate, precise and trusted **[2].**

Big Data comprehend all kind of data, but when we deal with data analysis or data projects is fundamental to comprehend what kind of data we are treating. It is fundamental to define the type of data because consequently, based on the type of data there will be a specific set of activities and task to perform on them. Typically, data are segmented in 3 main sets:

- **Structured Data**: they are data collected with a clear organization, usually with a tabular form where each event has different features characterizing it. Considering a table each row represents an event and each event has its own features that correspond to the number of columns.
- **Unstructured Data**: they are data collected without an organization, they are not clear and extracting information from them is very difficult because they are not pre-defined in a clear manner. When we talk about unstructured data, we refer to images, videos, records and so on.
- **Semi-structured Data**: they are data that belong to structured data but do not obey to formal structure, they may be not collected in a tabular form and they may not have all the same features.

Every day, every hour, every minute and every second many data are created by people, machine, companies, application and everything is digital and has the capability to record events and consequently data. In the last year we are seeing an increase of data generation due to the availability of interconnected devices to Internet, new software and cheaper solution to collect and store data.

A real impressive fact is that the 90% of the existing data were create only in the past 2 years **[3]**. Data are growing exponentially, and this is the reason why they are getting importance and value now, because before there were not so many data to extract information and patterns useful to conduct business decisions. As we can deduct from this situation, and how we can see from the market, nowadays the top-companies in the world are those ones that are able to use and extract very important information from data they create and gathered.

Big Data doesn't mean new hardware or software, it means new way of looking at data, new extended information outside the company and new type of analysis. To extract value from Big Data, we need analytical models. To create analytical models, we require a team of data scientists.

This rapid evolution of businesses, data creation and collection led to create a new method to manage and to leverage on data named Machine Learning, a method to deal with Big Data that enables computers to learn from data patterns and to develop models able to extract value from data with the supervision of high-skilled professionals, called data scientists. This professional figure will be the focus of the following Chapter 2.

1.3 Machine Learning Introduction

> *"Machine Learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence (AI). ML algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task."* **[4]**

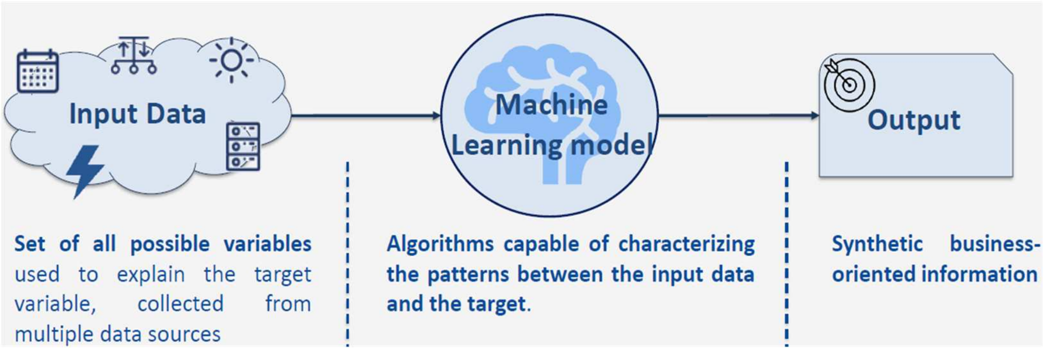

Input Data
Machine Learning model
Output
Set of all possible variables used to explain the target variable, collected from multiple data sources
Algorithms capable of characterizing the patterns between the input data and the target.
Synthetic business-oriented information

Figure 2 - How machine learning works

ML idea is to enable machines to learn how to behave from past data, allowing them to take actions or make predictions with a minimum human intervention. ML is characterized by different methodologies, techniques and tools on which perform its capabilities. It is possible to say that when we are talking about ML, we are intending different mechanisms that allow to an intelligent machine to improve its capabilities and performances during time.

At the base of ML there are a set of different algorithms that will be able to take one decision instead of another, or able to execute actions learned in the past. Before the birth of ML, the highest level of data analysis is what we call traditional statistics. To understand the changes brought by ML techniques is useful have a picture of what were the characteristics before and then ML:

Table 1 - Differences between traditional statistics and ML

	Emphasis	Modelling	Evolution	Generalizability
Traditional Statistics	Emphasizing on parameters interpretability, concerning over assumption and robustness	Preferring simpler models over complex ones, even if the performances drop, to be as general as possible	Focusing on a-priori hypothesis and statistical significance of results	Having statistical modeling or sampling assumption that connect data to a population of interest
Machine Learning	Emphasizing on prediction and performances over interpretability, concerning over robustness	Concern for overfitting but not model complexity per se, which can dynamically adapt to changes over time	Evaluating results via prediction performance (MAPE, MAE RMSE, Accuracy, Recall, Precision, etc..) on validation sets	Obtaining generalizability through performance on novel datasets (test sets – out of sample/time)

The two different solutions can be easily shaped with the two following schemes.

Traditional Statistics

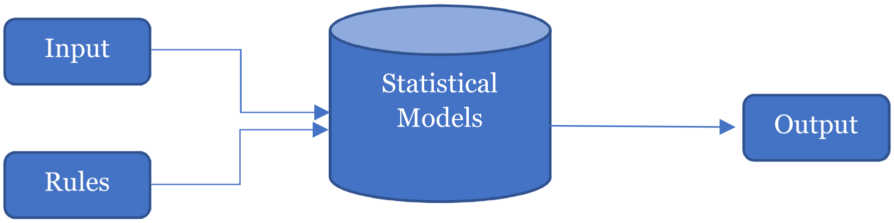


Figure 3 - Work-flow of statistical model development

Machine Learning

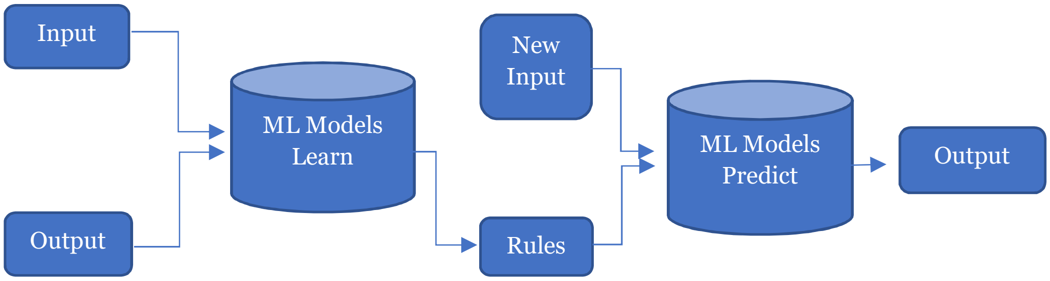


Figure 4 - Work-flow of ML model development

The big difference that we can immediately see is the conceptual method to approach the problem. If with the traditional statistics the objective is to find patterns inside data thanks to mathematics knowledge, with ML the objective is to allow machine to learn patterns from training data and then to apply these patterns (models) to new data never seen to make predictions or classification tasks. Statistics is at the base of ML but with a narrow purpose: analyze known data. Instead ML wide the purpose and tries to get accurate predictions thanks to the analysis of past data.

The algorithms of ML are different from each other for their approach to the problem, for the type of data input and output and the type of task or problem they are trying to solve. The algorithms can be divided as follow:

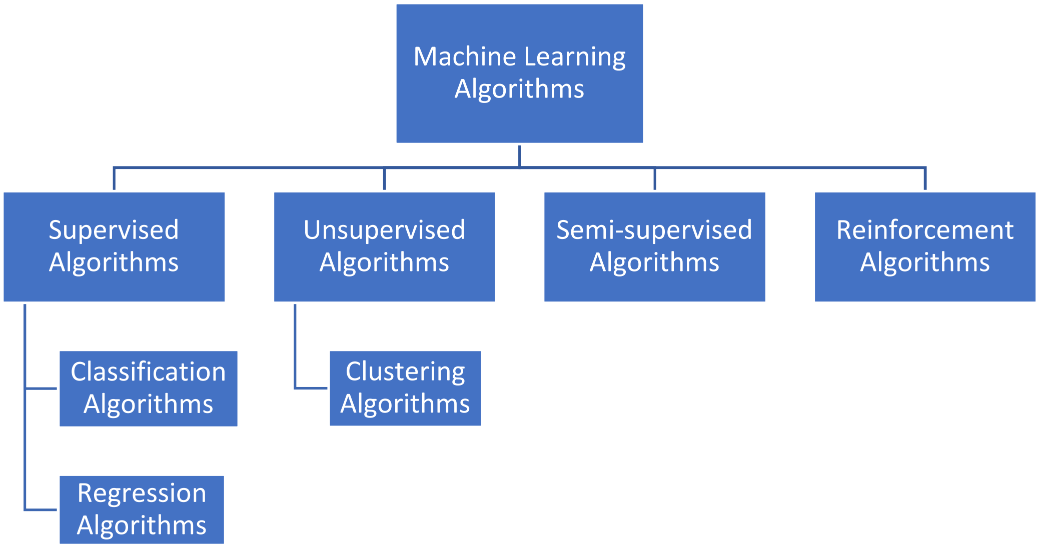


Figure 5 – Machine Learning algorithms

Supervised Learning

In this case, these algorithms build a mathematical model based on a set of data that contains both the inputs and the desired outputs. The data on which the model is build is called training data, and it is a dataset of records used to train the model. Supervised learning algorithms after the training phase are able to predict with high accuracy (it depends on the goodness of the model) the output of a new input set. Such algorithms are considered good when they are able to correctly determine the output for inputs that it has never seen before.

These algorithms could be both of classification or regression:

- Classification Algorithms: they are used when the output belongs to a restricted set of values.
- Regression Algorithms: they are used when the output may have any numerical value within a range.

Semi-supervised learning

It is similar supervised learning algorithms but in input data miss some outputs. The goal of a semi-supervised model is to classify some of the unlabeled data using the labeled information set.

Unsupervised Learning

In this case the algorithm takes a set of data in which are included just the inputs without any output. The algorithm has the task to find hidden patterns and structures in data that explain how data are correlated. The algorithm learns from input data that has not been labeled, categorized or classified. The most famous example is the cluster analysis on a dataset:

- Clustering Algorithms: they try to divide the observations of a dataset in subsets also called clusters, based on similarity or other predefined criteria. Consequently, if observations inside one cluster are similar, observations inside different clusters are dissimilar.

Reinforcement Learning

It is a branch of ML that is concerned with the computer programs or also called software agents, enable them take actions in an environment so as to maximize some notion of cumulative reward. It differs from supervised learning in that labelled input/output pairs need not be presented, and sub-optimal actions need not be explicitly corrected. Instead the focus is finding a balance between exploration and exploitation **[5]**.

1.4 Machine Learning Pipeline

A model that describe the organization of a ML project is the CRISP-DM framework. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It gives a structured planning even for a ML project. This model explains the state-of-the-art of a model development. In practice many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions.

The framework is composed by the following points:

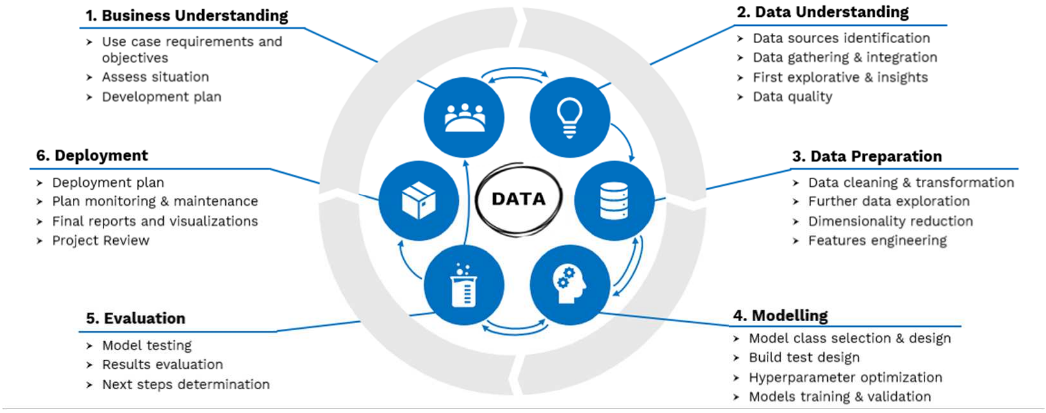

1. Business Understanding
Use case requirements and objectives
Assess situation
Development plan
2. Data Understanding
Data sources identification
Data gathering & integration
First explorative & insights
Data quality
3. Data Preparation
Data cleaning & transformation
Further data exploration
Dimensionality reduction
Features engineering
4. Modelling
Model class selection & design
Build test design
Hyperparameter optimization
Models training & validation
5. Evaluation
Model testing
Results evaluation
Next steps determination
6. Deployment
Deployment plan
Plan monitoring & maintenance
Final reports and visualizations
Project Review
DATA

Figure 6 - CRISP-DM Model

1. Business understanding

The first stage of the CRISP-DM framework is to understand the business objectives, uncover the hidden factors that could lead the project to better results. In order to define the right objectives to achieve is mandatory to assess the actual situation and then to develop a plan of actions.

2. Data understanding

The second stage of the model is centered on the data sources of the project. The important data must be identified, gathered and integrated. After having all important data is useful to do a first explorative analysis on them to extract some insights. Another important task to be done is to check the quality of data.

3. Data preparation

This objective of this stage is to prepare data before creating the ML model. Data to ingest in the model must be very relevant, data before being considered good has to pass different stage, they must be cleaned and transformed, they must be analyzed deeper to detect some important information. If only some data are relevant is necessary to apply some dimensionality reduction methods. If needed data can be created through some feature engineering techniques.

4. Modeling

The fourth phase of CRISP-DM model is the modeling, in this phase data scientists must select the ML model to apply and tune the hyper-parameters to improve model training. Once a model is trained it must be evaluated according some metrics chose to describe meaningfully the ability of the model.

5. Evaluation

Once a model is trained it must be evaluated according some metrics chose to describe meaningfully the ability of the model. This phase is characterized by tests applied to the model in order to understand if it fit well unknown situation.

6. Deployment

The last phase of the model is the implementation of the model, with the creation of a monitoring system to collect the results and keep the model always update if something goes wrong.

This model gives an overview of a traditional ML project, the next section will give you a deeper description of the different techniques involved in all the steps. Now that we have a general idea about the different tasks that a ML project must deal with, we can introduce some technical aspects to embrace when dealing with ML. A model, or better the model development, is the core activity of every ML project. The model is the logical structure developed to deal with specific data, aiming to obtain the expected output also from data never seen before. A model is a mathematical representation of a real-world process, it is derived from the training set by which the ML algorithm learn from. Creating a model is in general is a very complicated task, there are a lot of problems to manage and many steps to solve. Let's go to see the most famous and used techniques during the pipeline: the pipeline is the sequence of steps needed to build a consistent ML model.

The sequence of steps is not linear but iterative until the model is not good enough, as we have seen from the CRISP_DM organization could be present many backtrack between phases until the performance is not achieved. In each step the data are modified and transformed from a raw to a clean and understandable situation, depending on the algorithm to apply on the data.

Now let's go deeper in detail within each phase to understand the essential activities to perform on data. The objective of the following six subsections about the ML pipeline is to define the traditional activities performed during a project. The explanation of the main techniques involve in ML is useful to understand and to highlight the complexity involved behind an AutoML solution that typically works with a friendly drag and drop interface or with few steps where no code is necessary.

1.4.1 Data Cleaning

Once you have collected the raw data (raw data means data picked and collected by a source, they are set of data uncleaned and with possible errors and inconsistences), to become valuable for ML activities those data must be cleaned by the outliers, by the null values and so on. The first data preprocessing step of a ML pipeline is to clean your data.

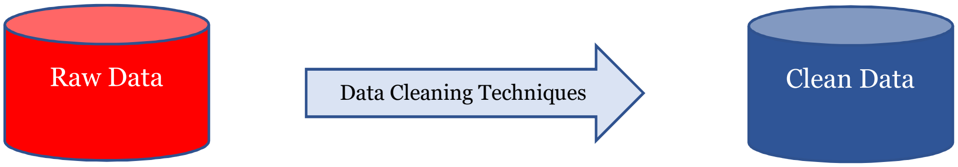


Figure 7 - First preprocessing step

We must know that data cleaning techniques vary from dataset to dataset based on the configuration of data and on the objectives of the analysis. Proper data cleaning can make or threat your project, it is a very important phase. Usually professional data scientists spend a very large portion of their time to complete in the better way possible this step.

Why is it so important? For two reasons well explained by the following two cites:

1. *"Garbage in is garbage out"* **[6]**

It means that if you build a ML model on bad data, even the model will result bad and the output won't be accurate as you want.

2. *"Better data beats fancier algorithms"* **[7]**

It means that it is better having a good dataset cleaned and preprocessed than a very good algorithm, the cause is that ML model are built on the data they are analyzing and the better they are the better will be the model.

Basically, both the "laws" say that in order to obtain a very good model, the first action to do is to transform skewed, raw and noisy data in a clean form suitable for algorithms to extract values from them.

The main methods to clean data are the following:

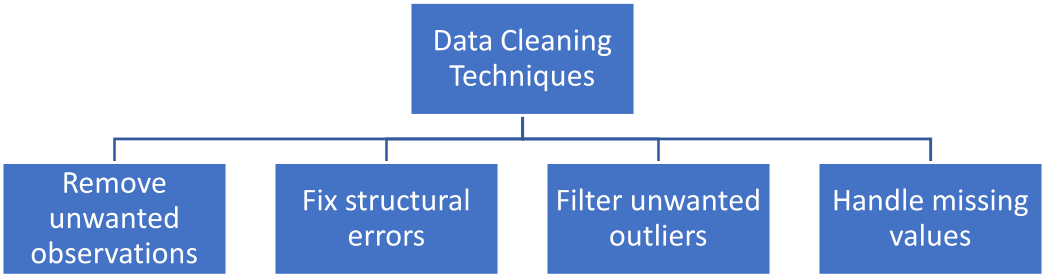


Figure 8 - Data cleaning main techniques

1. **Remove unwanted observations**, this includes duplicate or irrelevant observations.
2. **Fix structural errors**, these errors arise during measurement, data transfer, or other types of "poor housekeeping".
3. **Filter unwanted outliers**, they can cause problems with some models, but they are innocent until proven guilty, only if you have a legitimate reason to remove an outlier, it will help your model's performance.
4. **Handle missing data**, it is a deceptively tricky issue in applied ML because you can't just ignore them in your dataset since that many algorithms do not accept missing values. The two most commonly methods to handle missing values are dropping observations that have missing values or imputing the missing values based on other observations, but these two methods are not considered the more intelligent. Typically, you can have to deal with numerical of categorical missing values in a dataset and for each of them there are different techniques. With specific class of data as categorical ones, you can simply label them as "Missing", simply you are adding a new class and the algorithms can handle it. For Numeric data you should flag and fill the values, that means flag an observation with an indicator carriable of missingness and then fill the original missing value with 0 just to meet the technical requirement of no missing values.

The final objective of this phase is to have a high data-quality to feed the ML algorithms, high-quality data must pass a set of criteria to be considered 'good'.

- **Validity**: the degree to which the measures conform to defined business rules or constraints.
- **Accuracy**: the degree of conformity of a measure to a standard or a true value. It is very hard to achieve through data-cleansing in the general case, because it requires accessing an external source of data that contains the true value; such "gold standard" data is often unavailable.
- **Completeness**: the degree to which all required measures are known. Incompleteness is almost impossible to fix with data cleansing methodology: one cannot infer facts that were not captured when the data in question was initially recorded.
- **Consistency**: the degree to which a set of measures are equivalent in across systems. Inconsistency occurs when two data items in a dataset contradict each other. Fixing consistency is not always possible.
- **Uniformity**: the degree to which a set of data measures are specified using the same units of measure in all systems.

Once you are sure to have cleaned and made consistent your dataset, satisfying the precedent criteria, you can start building your ML algorithms.

We can notice that this part in the traditional way is very long and there are a lot of variables to consider. Usually this phase takes long time to be executed by data scientists. The next phase of the pipeline is a set of activities intercorrelated, it is the feature engineering phase.

1.4.2 Feature Engineering

It is the process in which typically data scientists use their domain knowledge of the data to create features that allow the best application of ML algorithms. It is a fundamental phase of ML, and it is both difficult and expensive in time and effort. The objective of this phase is to create and select the best set of features to build the model.

> *"Coming up with features is difficult, time-consuming, requires expert knowledge. 'Applied machine learning' is basically feature engineering."*
> (Andrew Ng, Stanford University)

A feature is an attribute or property shared by all the independent units on which analysis or prediction is to be done. Any attribute could be a feature, until it is useful to the model. The purpose of features, other than being attributes, would be much easier to understand in the context of a problem. Feature are important for predictive models because they influence the results that you are going to achieve with the model selected. The quality and quantity of the features will have great impacts on whether the model is good or bad. (in our business case in Chapter 4 will be faced a predictive problem solved with different models built on datasets with different set of features, with the objective to understand how the results rely on features in input.

Choose the right features is very important because better features can create simpler and more flexible models, and they often yield better results.

The typical process of feature engineering is composed as follow:

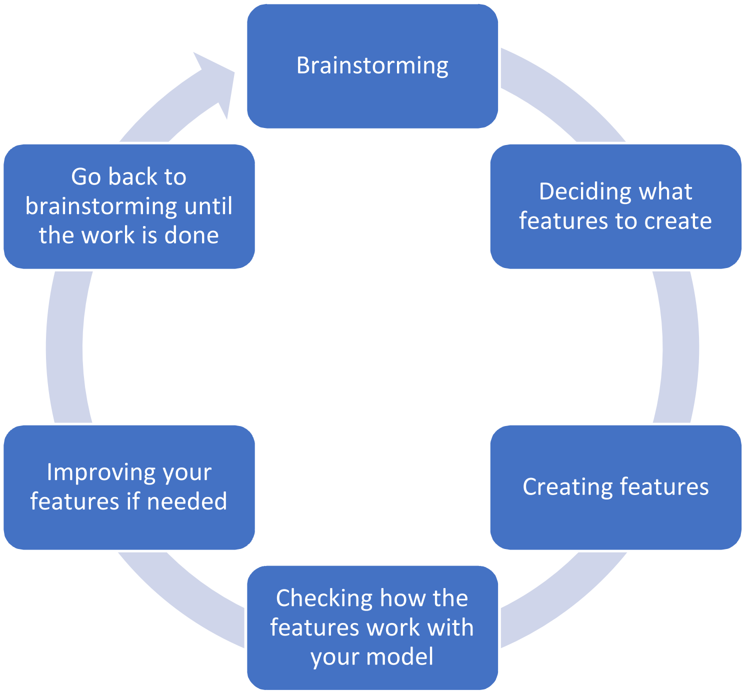

Brainstorming
Deciding what features to create
Creating features
Checking how the features work with your model
Improving your features if needed
Go back to brainstorming until the work is done

Figure 9 - Feature engineering steps

Typically feature engineering is iterative and based on the try & error approach, indeed the approach model shape is circular to permit iteration and improvement of results. Now we will introduce what are the main techniques that could be implemented to perform feature engineering tasks **[8]**:

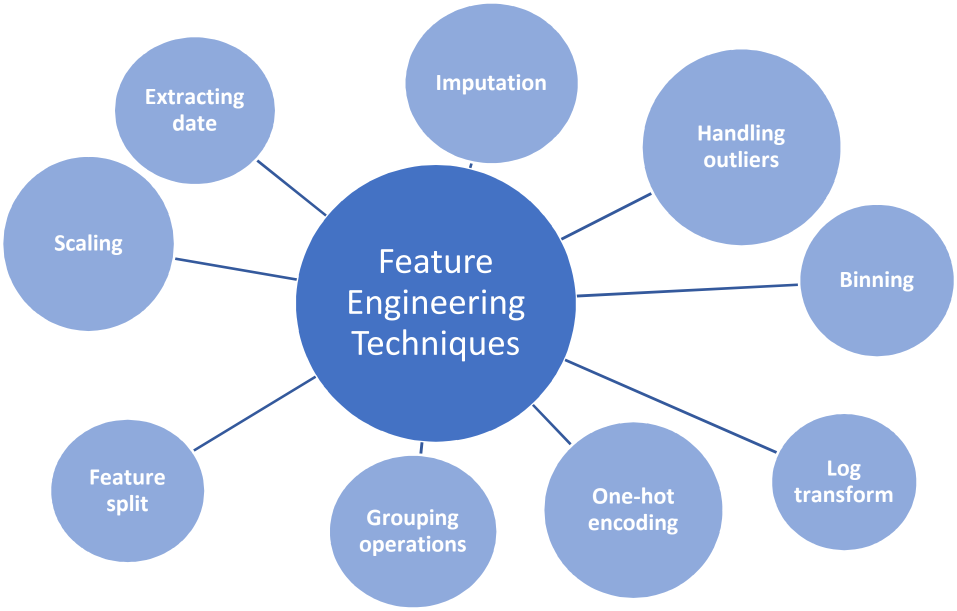


Figure 10 - Main feature engineering techniques

Imputation

It is better than dropping rows with missing values because it preserves the data size. For numerical values you can change NA with ‘0’ in column Boolean. Otherwise could be correct substitute missing values with the medians of the column. For categorical values instead replacing the missing values with the maximum occurred value in a column is a good option.

Handling Outliers

Outliers can be detected with standard deviation, if a value has a distance to the average higher than *x*standard deviation*, it can be assumed as an outlier.

Outlier detection with percentiles, it is another mathematical method. You can assume a certain percent of the value from the top or the bottom as an outlier, the key point here is to set the percentage value once again, and it depends on the distribution of your data.

Binning

Binning can be applied both to numerical and categorical data, the main objective of binning is to make the model more robust and prevent overfitting, it has a cost to the performance. The trade-off between performances and overfitting is the key point of the binning process.

Table 2 - Example of binning technique, on the right for numerical features and on the left for categorical features

Numerical binning example	
Value	Bin
0-30	Low
31-70	Mid
71-100	High

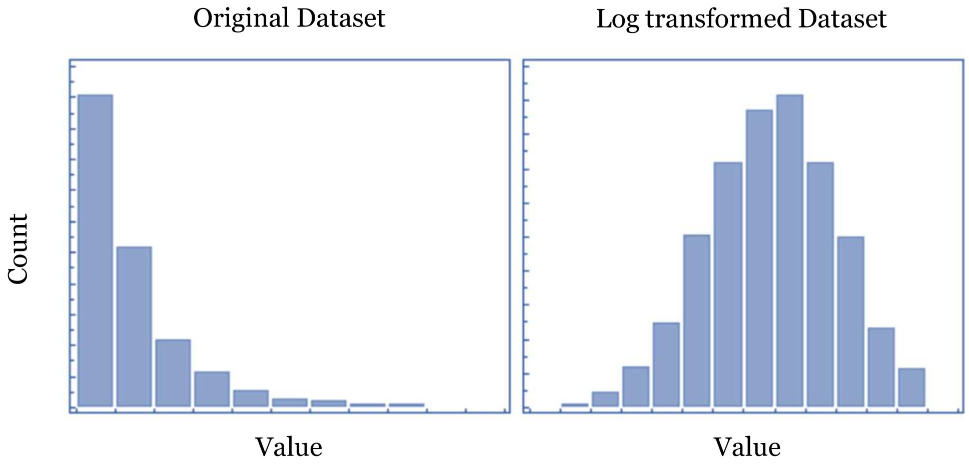
Categorical binning example	
Value	Bin
France	Europe
Italy	Europe
Brasil	South America

Logarithmic Transformation

Logarithmic transformation is one of the most commonly used mathematical transformation in feature engineering. It has different benefits:

- It helps to handle skewed data, after the transformation the distribution becomes more approximate to normal.
- In most of the cases the magnitude order of the data changes within the range of the data.
- It also decreases the effect of the outliers, due to the normalization of magnitude differences and the model become more robust.

Figure 11 - Example of log transformation



One-hot encoding

It is one of the most common encoding methods in ML. This method spreads the values in a column to multiple flag columns and assign 0 or 1 to them. These binary values express the relationship between grouped and encoded column. This method changes your categorical data, which is challenging to understand for algorithms, to a numerical format and enables you to group your categorical data without losing any information.

> *"Why One-Hot? If you have N distinct values in the column, it is enough to map them to N-1 binary columns, because the missing value can be deducted from other columns. If all the columns in our hand are equal to 0, the missing value must be equal to 1. This is the reason why it is called as one-hot encoding"* **[8]**

Table 3 - Example of one-hot encoding

UserID	City
UID76	Roma
UID32	Madrid
UID45	Madrid
UID09	Instanbul
UID33	Instanbul
UID12	Instanbul
UID75	Roma

UserID	Instanbul	Madrid
UID76	0	0
UID32	0	1
UID45	0	1
UID09	1	0
UID33	1	0
UID12	1	0
UID75	0	0

Grouping operations

In most ML algorithms, every instance is represented by a row in the training dataset, where every column shows a different feature of the instance. This kind of data are called "tidy". There are 3 main ways to aggregate categorical columns:

- First option to select the label with the highest frequency.
- Second option is to make a pivot table. Instead of binary notation, it can be defined as aggregated functions for the values between grouped and encoded columns.
- The third and last option is to apply a group by function after applying one-hot encoding. This method preserves all the data, and in addition, you transform the encoded column from categorical to numerical in the meantime.

Numerical columns are grouped using sum and mean functions in most of the cases. Both can be preferable according with the meaning of the feature.

Feature split

Splitting feature is a good way to make them useful in terms of ML because usually datasets contain string columns that violates tidy data principles. By extracting the utilizable part of a column into new feature:

- ML algorithms able to understand them.
- Make possible to bin and group them.
- Improve model performance by uncovering potential information.
- There is not a single way to split feature, it depends on the feature itself.

Scaling

Many times, the numerical features of the dataset do not have a certain range and they differ from each other. How can ML comprehend for example the range of two columns different as '*Age*' and '*Income*', how these two columns can be compared? Scaling solve this problem, the continuous features become identical in terms of the range, after a scaling process. Algorithms based on distance calculations such as k-NN or K-Means need to have scaled continuous feature as model input. Two main ways:

- Normalization, it scales all values in a fixed range between 0 and 1. It does not change the distribution of the features and due to the decreased std dev, the effects of the outliers increases. It is recommended before applying scaling to handle the outliers with care

$$Xnorm = \frac{X - Xmin}{Xmax - Xmin}$$

- Standardization scales the values and at the same time takes in consideration standard deviation. If the standard deviation of features is different, their range also would differ from each other. This reduce the effect of the outliers in the features.

$$z = \frac{x - \mu}{\sigma}$$

Extracting date

Through date columns usually provide valuable info about the model target, they are neglected as an input or used nonsensically for the ML algorithms. Building an ordinal relationship between the values is very challenging for a ML algorithm if you leave the date columns without manipulation. Three possible types of preprocessing for dates:

1. Extracting the parts of the date into different columns: Year, month, day, etc.
2. Extracting the time periods between the current date and columns in terms of years, months, days, etc.
3. Extracting some specific features from the date: Name of the weekday, weekend or not, holiday or not, etc.

1.4.3 Feature selection

It is the process of selecting a subset of relevant features to use in model construction.

Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, whereas feature selection methods include and exclude attributes present in the data without changing them.

When we deal with feature selection the objective is to choose and apply the selected features to our model to see if it works good or not. It is a fundamental part because after having cleaned and engineered our raw data if we don't choose the essential features to build our model all the effort made will be lost, since the model will not work as expected.

Feature selection is one of the core concepts in ML which hugely impacts the performance of your model. The data features that you use to train your models have a huge influence on the performance you achieve. If on one side some features can benefit your model, on the other side some of them can affect negatively its performances.

Feature selection techniques are used for four main reasons:

- To short training times.
- To avoid the curse of dimensionality.
- To enhance generalization by reducing overfitting.
- Simplification of models to make them easier to interpret by researchers.

Now we are going to discuss the various techniques and methodologies that you can use to subset your feature space and help your models perform better and efficiently **[9]**.

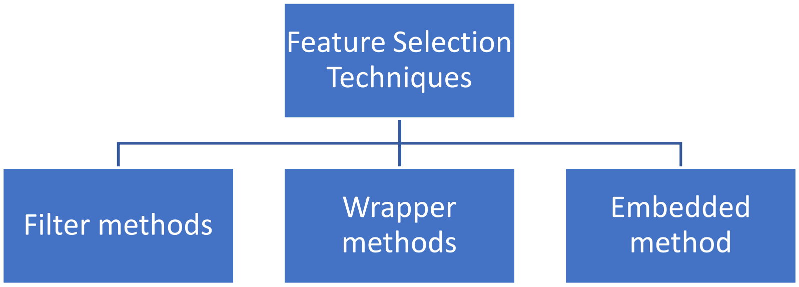


Figure 12 - Feature selection techniques

Filter methods

Features are selected based on their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here.

Wrapper methods

With this method we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is reduced to a search problem. These methods are usually computationally very expensive. Some common examples of wrapper methods are forward feature selection, backward feature elimination and recursive feature elimination.

Embedded methods

These methods combine the qualities of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods. Some of the most popular methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting.

Modeling is the next phase of the traditional pipeline of a ML project. After having changed all the raw data in clean data, after having transformed and created relevant features and chose the best ones, now is time to introduce the phases of model selection and hyper-parameters optimization. In the following two section we will discuss about these steps and what are the main activities to perform in order to create a good model.

1.4.4 Model selection

> *"In model selection tasks, we try to find the right balance between approximation and estimation errors. More generally, if our learning algorithm fails to find a predictor with a small risk, it is important to understand whether we suffer from overfitting or underfitting."* **[10]**

When dealing with model selection, the attention goes to avoid two main problems:

Overfitting: in the field of ML happens when an algorithm is fitted with a certain set of examples called training test. Typically happens when we have a set of data on which we know the results and another set on which we want to predict the future results. The algorithm will achieve a level of learning that enable it to predict the set not analyzed yet. But when the fitting phase is too long or the training set is too small, the model shall adapt to characteristics unique of the training set, but that are useless for the test sets. So, the model will be very accurate for the training set but will give low performances on the test set. From the example is clear that the blue line shows overfitting characteristics with an irregular model to determine value, instead with the green line we have a stable model even with data never seen before.

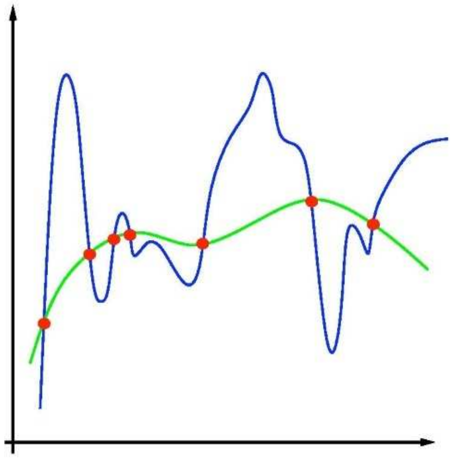


Figure 13 – Overfitting, blue line

Underfitting: it occurs when a statistical model cannot adequately capture the underlying structure of the data. A model is underfitted when some parameters or terms that would appear in a correctly specified model are missing. For instance, underfitting would occur when fitting linear model to a non-linear data. Such a model will tend to have poor predictive performance. The red line in figure shows a model than is underfitted, we can see that it does not fit the red points as well as the green line which represents a good model that is not characterized neither by underfitting nor overfitting.

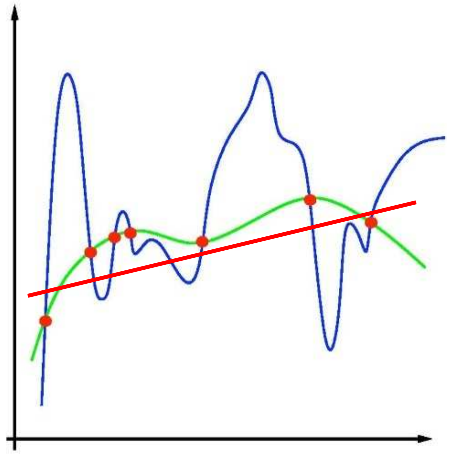


Figure 14 – Underfitting, red line

When we build a model, we must check if the model is getting overfitting or underfitting, how we can do this? To address this, we can split the original dataset into separate training and test subsets.

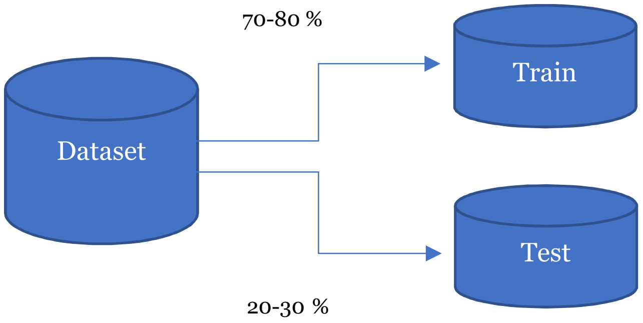


Figure 15 - Dataset splitting methods: train and test

This method can approximate of how well our model will perform on new data. If our model does much better on the training set than on the test set, then we are likely overfitting.

Avoid these two problems is very important to build a correct model that have optimal performances on data never seen. It is a part where the expertise of the data scientist is very crucial because to prevent these two problems, he/she can adopt different strategies, the most commons are:

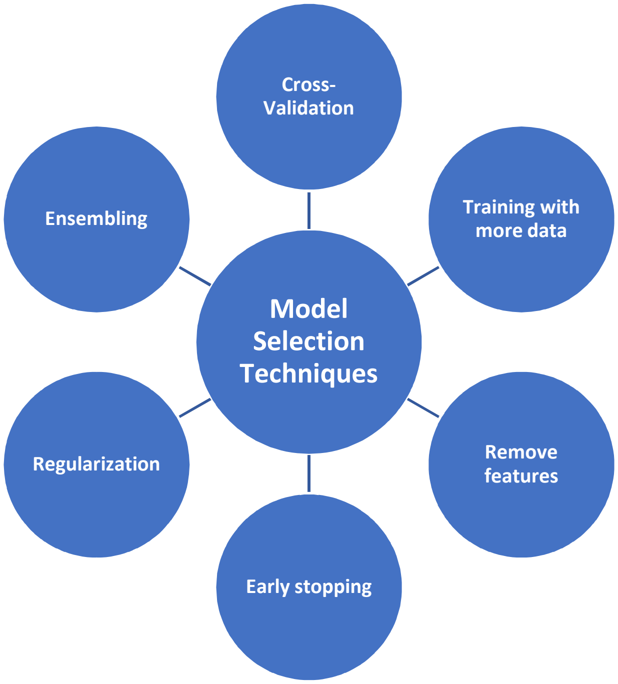


Figure 16 - Model selection techniques

Cross-Validation

It is a powerful preventative measure against overfitting. The idea is to use your initial training data to generate multiple mini train-test splits and use these splits to train your model.

In standard k-fold cross-validation, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining fold as the test set called holdout fold.

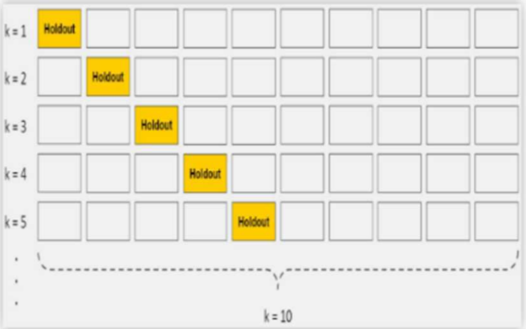


Figure 17 - Cross validation technique

Train with more data

It doesn't work every time, but training with more data can help algorithms detect the essential features. Obvious that if we add more noisy data, this technique is useless. So before to add more data we should be always ensure that data are clean and relevant.

Remove features

Some algorithms have built-in feature selection, for those that don't, we can manually improve their generalizability by removing irrelevant input features. It is a manual task that request time and deep knowledge of the data you are considering. It can even apply this technique with a try and error approach, testing which are the features more important with different subsets.

Early stopping

When we are training a learning algorithm iteratively, you can measure how well each iteration of the model performs. Up until a certain number of iterations, new iterations improve the model. After that point the model's ability to generalize can weaken as it begins to overfit the training data.

Early stopping refers stopping the training process before the learner passes that point. It is a technique to avoid the overfitting.

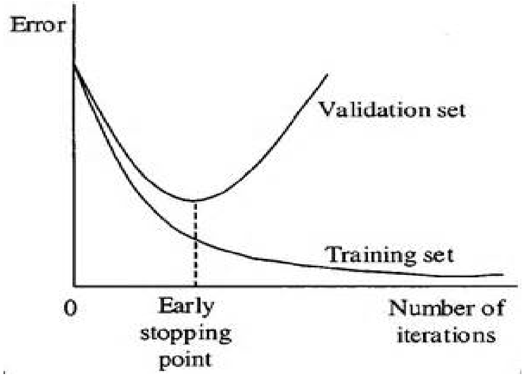


Figure 18 - Early stopping technique

Regularization

It refers to a broad range of techniques for artificially forcing your model to be simpler. The method depends on the type of learner (algorithm) you are using. For example, you could prune a decision tree, use dropout on a neural network, or add a penalty parameter to the cost function in regression.

Ensembling

Ensembles are ML methods for combining predictions from multiple separate models. There are two main methods for ensembling:

Bagging:

- It attempts to reduce the chance overfitting complex models.
- It trains large number of 'strong' learners in parallel.
- A strong learner is a model that is relatively unconstrained.
- Bagging then combines all the strong learners together in order to 'smooth out' their predictions.

Boosting:

- It attempts to improve the predictive flexibility of simple models.
- It trains large number of 'weak' learners in sequence.
- A weak learner is a constrained model.
- Each one in the sequence focuses on learning from the mistakes of the one before it.
- Boosting then combines all the weak learners into a single strong learner.

They are both ensemble methods but their approach to the problem is from opposite directions.

The next and last phase of the pipeline we consider is the hyper-parameters optimization. It is a very hard task aimed to improve the performances of the model selected.

1.4.5 Hyper-parameters optimization

The purpose of this step is to improve the learning process of the algorithm used to train the model. In the world of ML there are two types of parameters:

1. **Model parameters**
   They are learned by the algorithm while learning phase.
2. **Hyper-parameters**
   A hyper-parameter is a parameter whose value is used to control the learning process. They need to be set before beginning the learning phase.

Optimizing the hyper-parameters is a function with the objective of minimizing the loss/cost of the algorithm, which in turn keep balance between the mode bias and variance. This is essential in getting a low cross-validation error at the end of the experiment. There are different techniques to tune hyper-parameters:

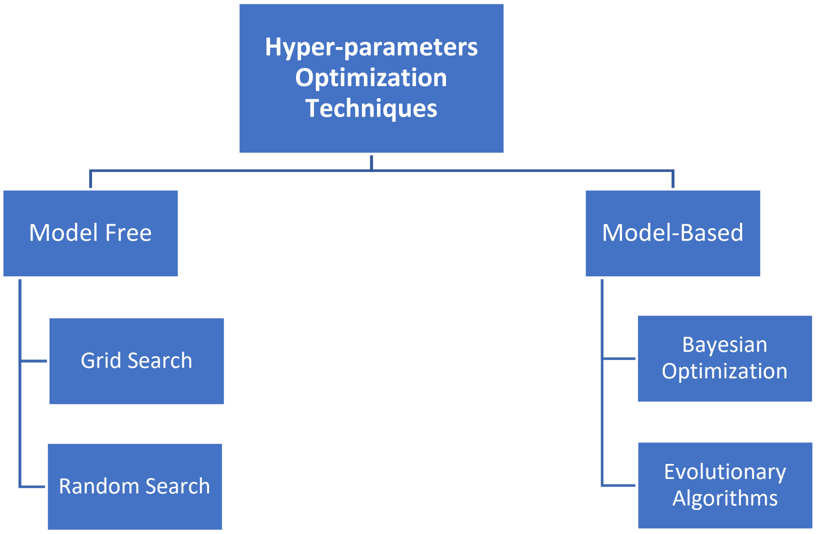


Figure 19 - Hyper-parameters optimization techniques

The objective of this part is not to explain all the possible techniques in detail but to give an overview of the main techniques to adopt. The more important techniques are:

Grid Search

Grid search expects few sets of values as parameter space and tries all combinations of these values to learn in brute force manner. Search will be guided by a metric, which is often cross validation error of the training data or evaluation on the test data.

Grid Search suffers from curse of dimensionality, because even when there are two hyper-parameters and five distinct values of these parameters, it requires twenty-five times of modeling and evaluation. Besides, there is no feedback or adjustment mechanism, thus the algorithm is highly unintelligent.

Random Search

Random search is very similar to grid search and does pretty much the same, but in a random combination of hyper-parameters. It is proved to outperform Grid Search, but it performs poorly in real cases as there is not adjustment or feedback in the learning process based on the results of previous learning.

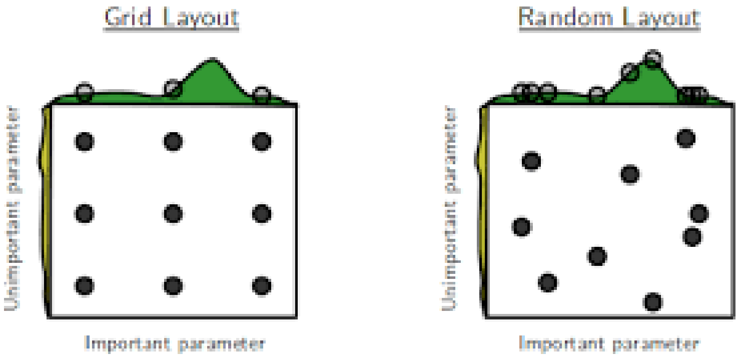


Figure 20 – Grid and Random search

Bayesian Optimization

Bayesian optimization is a global optimization technique for noisy black-box functions. Applied to hyperparameter optimization, Bayesian optimization shapes a probabilistic model of the function mapping from hyperparameter values to the objective evaluated on a validation set. By iteratively evaluating a promising hyperparameter configuration based on the current model, and then updating it, Bayesian optimization, aims to gather observations revealing as much information as possible about this function and the location of the optimum. It tries to balance exploration (hyperparameters for which the outcome is most uncertain) and exploitation (hyperparameters expected close to the optimum) **[11]**.

Evolutionary Algorithms

Evolutionary optimization is a technique for the global optimization of noisy black-box functions. In hyperparameter optimization, evolutionary optimization uses evolutionary algorithms to search the space of hyperparameters for a given algorithm. Evolutionary hyperparameter optimization follows a process inspired by the biological concept of evolution **[11]**:

1. Create an initial population of random solutions.
2. Evaluate the hyperparameters tuples and acquire their fitness function.
3. Rank the hyperparameter tuples by their relative fitness.
4. Replace the worst-performing hyperparameter tuples with new hyperparameter tuples generated through crossover and mutation.
5. Repeat steps 2-4 until satisfactory algorithm performance is reached or algorithm performance is no longer improving.

What we have explained until now are the general approaches to do when developing a ML model. The only things that remain to define are the metrics used to assess the goodness of a model. The next section will introduce the main evaluation methods for ML models.

1.4.6 Model evaluation

Once the model is created, the phase of testing includes the evaluation of the model based on certain metrics. Metrics are measures to define the goodness of the model according to the problem it is solving. Basically, when talking about evaluation metrics is savvy to separate the metrics used to evaluate classification models from regression models because the scopes are different and consequently even the metrics have different meanings. To have a comprehensive view of these metrics we are going to list the most important ones both for classification models and for regression models.

Classification metrics

- Confusion matrix
  It is one of the easiest and most intuitive metrics to assess the correctness of a classification model. To understand this metric must be known some concept:

  - TP (True Positive), TP are the cases when the actual class of the data point was True and the predicted is also True.
  - FP (False Positive), FP are the cases when the actual class of the data point was False and the predicted is True.
  - TN (True Negative), TN are the cases when the actual class of the data point was False and the predicted is False.
  - FN (False Negative), FN are the cases when the actual class of the data point was False and the predicted is True.

		Actual Value	
		positive	negative
Predicted Value	positive	TP	FP
	negative	FN	TN

Table 4 - Confusion Matrix

The objective is to reduce the more possible the FP and FN, while increasing the number of TP and TN. Once you defined these classes with the predictions of the model you implemented, then you can measure other important metrics that better describe the results of the model.

- **Accuracy**
  The accuracy of a model refers to the number of good predictions over all the predictions made. It can be calculated thanks to the confusion matrix:
  $$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Precision**
  Precision is a measure that tells us what proportion of data points that we diagnosed as positives, and actually were positive.
  $$Precision = \frac{TP}{TP + FP}$$

- **Recall**
  Recall is a measure that tells us what proportion of data points that actually are positive were diagnosed by the algorithm as positive.
  $$Recall = \frac{TP}{TP + FN}$$

- **F1 Score**
  It is the weighted average between Precision and Recall. It is not intuitive as precision or recall metrics but F1 is more useful than Accuracy, mainly if you are considering an uneven number of classes to predict.

$$F1\ score = 2 * \frac{Precision * Recall}{(Precision + Recall)}$$

- **ROC curve**
  It is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. It is created by two dimensions: TPR (True Positive Rate) and FPR (False Positive Rate) where:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

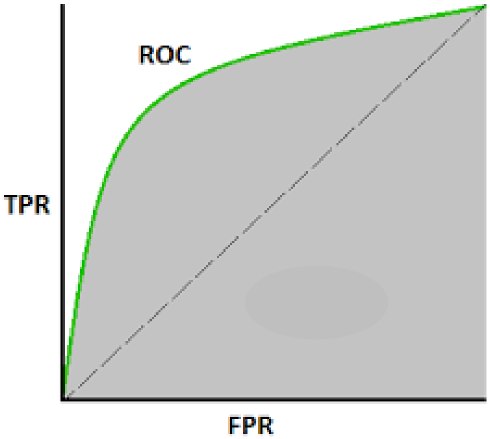


Figure 21 - ROC curve

- **AUC (Area Under the ROC Curve)**
  AUC provides an aggregate measure of performance across all possible classification thresholds. AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

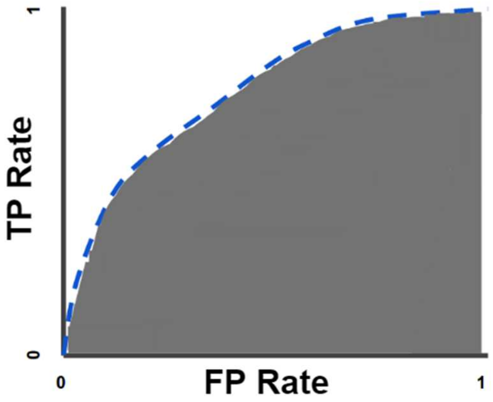


Figure 22 - AUC

Regression metrics

- **MAE (Mean Absolute Error)**
It is the simplest regression metric to understand. It is the absolute average value of the residual between predicted and actual value.

$$MAE = \frac{1}{n}\sum_{1}^{n}|Prediction - Actual|$$

Graphically the MAE can be explained as follow in the next picture:

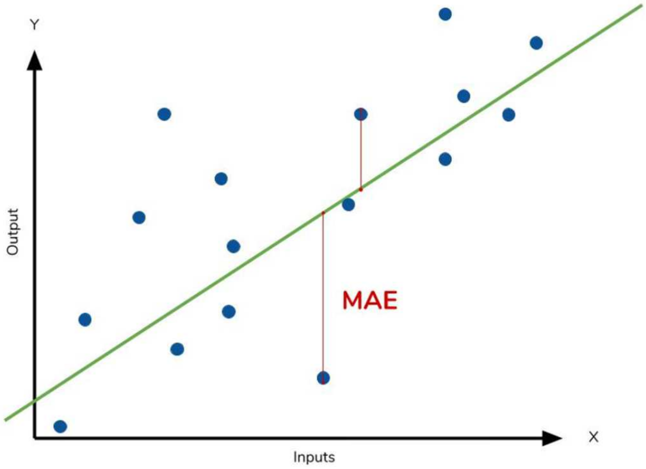


Figure 23 - MAE

- **MAPE (Mean Absolute Percentage Error**)
It is the relative of the MAE but in percentage.

$$\text{MAPE} = \frac{100\%}{n}\sum_{1}^{n}|\frac{Predicted - Actual}{\text{Actual}}|$$

As we will see in Chapter 4, MAPE in our tests is very important and it will be the metric adopted to assess the goodness of the AutoML models that had been created to solve our business problem.

- **MSE (Mean Square Error)**
It is the square of the MAE. It does not manage well situation in which there are outliers.

$$MAE = \frac{1}{n}\sum_{1}^{n}|Prediction - Actual|^2$$

- **RMSE (Root Mean Squared Error)**

$$RMSE = \sqrt{\frac{\sum_{1}^{n} |Prediction - Actual|^2}{n}}$$

It is one of the most used metrics to assess the goodness of a regression model, it is insightful because it displays the plausible magnitude of error term. It is highly affected by the presence of outliers, so before to be applied you must eliminate all the outliers by your data.

The metrics introduced above are the most used in classification and regression models evaluation. There are many others important metrics to consider but for the purpose of the thesis the metrics cited are enough. For our experiments we used one specific metric, the MAPE, because in our business case it is directly related to an economic value and so it is considered as the main KPI to consider when we evaluated our AutoML models, these concepts will be explained deeply in chapter 4.

Before going ahead let's have a summary of what we seen until now. We introduced the aim of ML to create value from data, the traditional organization of a ML project and all the essential phases of the pipeline from data preprocessing to the model evaluation to understand the potential impact that AutoML could bring in the ML lifecycle.

As we know the complexity is very high and each model is different from each other in terms of objectives, context and data considered. The next section will discuss the main problems in ML that lead us to introduce the core topic of the thesis: the automated machine learning as an alternative to ML and as an innovation in the world of data science.

1.5 Machine Learning problems

How we seen in the previous paragraph, when we deal with a ML model development there are a lot of things to consider and to check in every step. Many tasks are very time and effort consuming for data scientists, especially the data preparation phase which include collecting data, cleaning data, engineering data and selecting features. From a survey launched by CrowdFlower in 2017 on a sample of 80 data scientist emerged that the 80% of their time is used for data preparation (cleaning data and collecting datasets).

The next pie chart shows the result of this survey **[12]**:

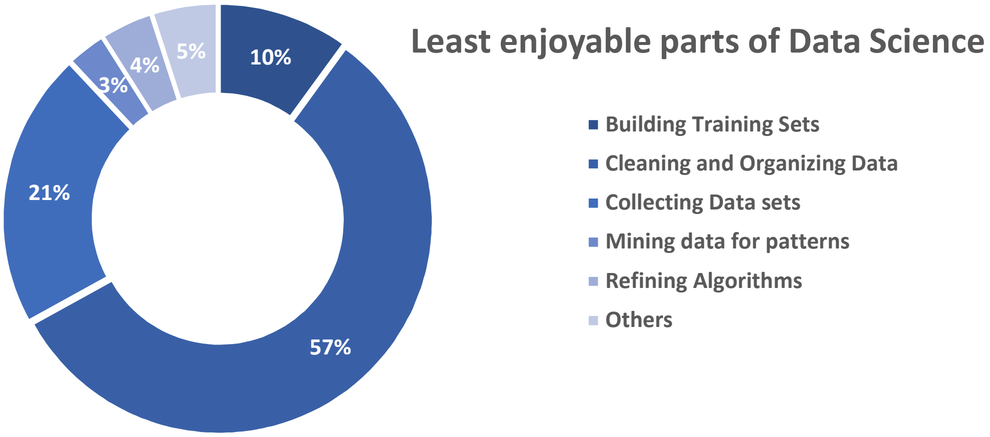


Figure 24 - Results of the survey made by Forbes

The chart highlights that only 20% of data scientists' time is allocated to create machine learning models or mining data patterns. Instead the rest of the time is used to make tedious and repetitive tasks to let usable data skewed and noisy. From this survey emerged that 76% of data scientists view data preparation as the least enjoyable part of their works.

Another unfriendly task to solve is the hyper-parameters optimization in which data scientists try many combinations of hyper-parameters to find the best configuration of the algorithm that fit with the problem.

The main problem is that there are many tasks that are repetitive and take a lot of time to be solved because usually humans commit biases and errors. It is here that finally we can introduce the concept of automated ML. AutoML was born with the objective to make easier the ML tasks, automating totally or at least on part of the traditional pipeline seen before. AutoML aims to decrease the time and effort spent by data scientists in doing repetitive and time-consuming tasks, leaving to them the added-value activities in which their expertise is fundamental.

The second big problem of ML is that just few very specialized people are able to develop and to deploy ML model, and the amount of professional is not appropriate to the demand of the market today. The consequence is that it is very difficult to find professionals. What we learned about AutoML is that it has the main objective to automate ML, but it introduces also some consequences for the professional roles who adopt it. The main observation is that AutoML enables also people who do not have deep knowledge in data science to deal with ML projects and to manage data in a real high-value method. This leads the market of data science to several possible future scenarios in which there are many variables to consider like the arise of new professional roles called citizen data scientists, new competencies needed to use AutoML and to program it, a possible boom of data-driven approaches in industries where data are not used to take strategic decisions. All these aspects will be analyzed in the chapter 2.

In order to compare the new AutoML solutions with the past solutions, let's fix the benefits and constraints of the traditional ML before introducing AutoML:

Table 5 – Benefits and constraints of ML

Benefits	Constraints
• High value analysis of data • Very accurate results • Customized solutions • Identification of hidden patterns in data	• Long time and high effort to perform different steps of the pipeline • Expensive projects • Need organizations of data science team • Very skilled individuals • Human biases • Manual errors

1.6 Automated Machine Learning

There is not a formal definition for automated machine learning. AutoML is an 'umbrella' term coined from '*automated machine learning*', it refers to a large-scale automation of a wide spectrum of ML processes. (ChaLearn AutoML challenge, 2015)

AutoML gained visibility and popularity after the ChaLearn Automated Machine Learning competition initiated in 2015. It was started as a benchmark for AutoML systems that can be operated without any human intervention, the challenge focused on hyper-parameters tuning and model selection for classification learnings.

The field of AutoML aims to make decisions in a data-driven, objective, and automated way: the user simply provides data, and the AutoML system automatically determines the approach that performs best for a particular application. Thereby, AutoML makes state-of-the-art ML approaches accessible to domain scientists who are interested in applying ML but do not have the resources to learn about the technologies behind it in detail. This should be considered as a democratization of ML: with AutoML, customized state-of-the-art ML is at everyone's fingertips. (Frank Hutter et al., AutoML Book, 2018)

A recent Google Research article explains that:

> *"The goal of automating ML is to develop techniques for computers to solve new machine-learning problems automatically, without the need for human-ML experts to intervene on every new problem. If we're ever going to have truly intelligent systems, this is a fundamental capability that we will need."* **[13]**

AutoML provides methods and processes to make ML available for non-ML experts, to improve efficiency of ML and to accelerate research on it. AutoML has achieved considerable successes in recent years and an ever-growing number of disciplines rely on it. The main reasons why AutoML is used nowadays are:

- Preprocess the data
- Manage and select appropriate features
- Select an appropriate model family
- Optimize model hyper-parameters
- Model evaluation

As the complexity of these tasks is often beyond non-ML experts, the rapid growth of ML applications has created a demand for off-the-shelf ML method that can be used easily and without expert knowledge. As a new sub-area in AI, AutoML has got more attention not only in ML but also in computer vision, natural language, processing and graph computing (example of Google AutoML Vision and Clarifai).

Theoretically AutoML can lead to improved performances while saving substantial amounts of time and money, as ML experts are both hard to find and expensive. As a result, commercial interest in AutoML has grown dramatically in recent years, and several major tech companies are now developing their own AutoML systems.

To make easier the AutoML concept, a powerful explanation is given by representing the process of a machine learning model development with and without AutoML, in this way is understandable how it works and when it can be used. The figure is taken by the documentation of TPOT **[14]**, it is a good image of the traditional pipeline of machine learning:

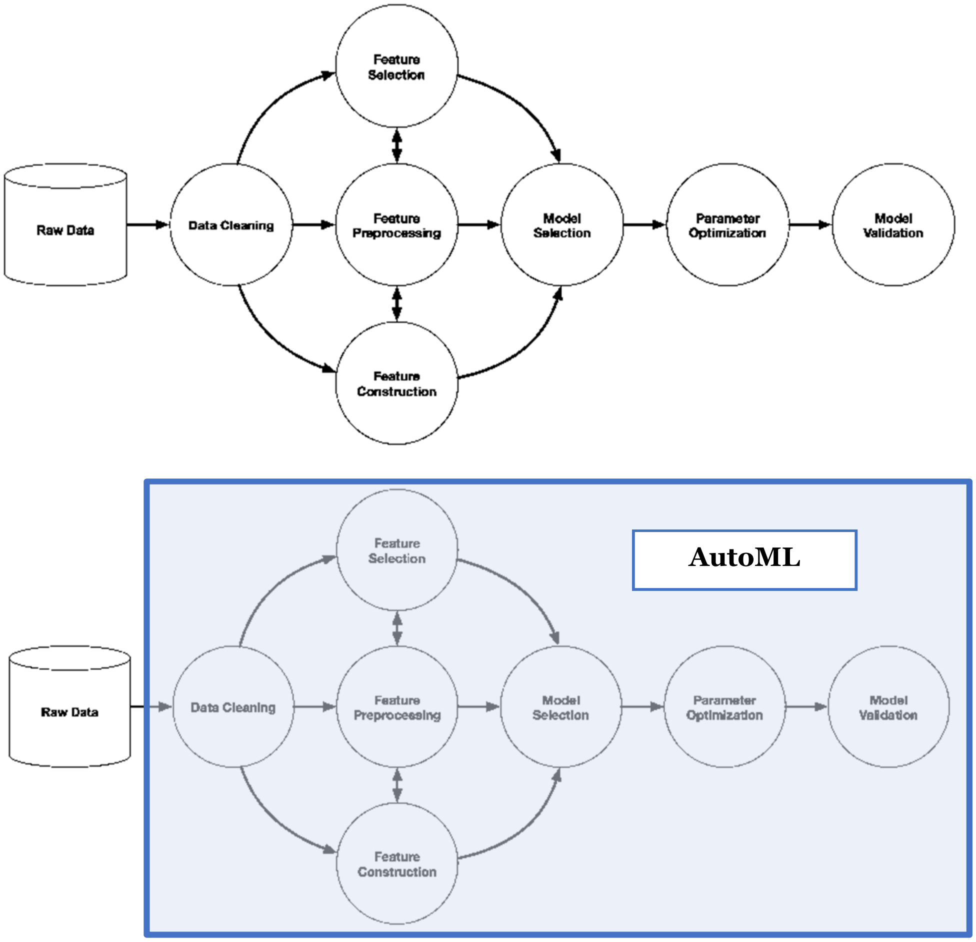


Figure 25 - Graphical pipeline with and without AutoML

Conceptually AutoML is the automation of the entire pipeline of a machine learning model, once you have uploaded your data you can run the system and it executes all the phases without human intervention. This is a general description of AutoML because we will see that not every system automates the entire pipeline but only a specific phase.

Furthermore, about AutoML there are many discussions about the impact that it could bring to the field of data science, mainly under the perspective of the professional role of data scientist and the future of data science team composition. What emerged are suppositions whether the data scientist could gain benefits or constraints from this new technology.

In the next and last section of the introduction chapter we will define the overall benefits identified related to AutoML, we are going to define even the theoretical constraints.

1.7 AutoML Benefits and Constraints

In this chapter we are going to define and motivate the benefits that AutoML ideally can bring and at the same time what are its main constraints and limits.

Benefits	
Increase productivity of data scientists	Reducing time needed for tedious and repetitive tasks as data cleansing and hyper-parameters optimization, allowing them to focalize their effort on real value-added tasks.
Ease of usage	Many AutoML system are free-code, that means that the program works just without any line of code. AutoML enables even people without strong knowledge in data science or without knowledge in programming language as Python or R to create models.
Fill ML experts demand	Today the need of data science professional figures is imbalanced with their availability. With AutoML this gap could be filled since many data science tasks can be executed autonomously by the AutoML systems exploited by 'citizen data scientist'. We will define who is the citizen data scientist in the Chapter 2.3.
Create new job position	Not only data scientist can work on data, now also people without a specific background can extract important values from data, performing analysis, developing ML models.
Democratization of ML	ML is adopted only in companies who have knowledge in managing data and the resources to do that. Today the number of companies that leverage their business decisions on data are increasing and now, with AutoML, everyone can adopt ML to run its business without an expensive data science team to hire.
Errors reduction	AutoML reduces bias and errors that occur when a human being is designing the ML models. The possible errors came from the bad designing of the AutoML system itself.

Table 6 - AutoML benefits

Constraints	
Reduction of ML knowledge	Since that companies can use AutoML tools, employees do not need to learn mathematic, language programming and statistics because it is all automatic. Instead they need to learn how to use the new technologies and how to interpret results
Complex data issues	Many autoML tools work well when they deal with simple data in a clear context. If data are complex or un-structured the tool may fail in adapting ML model to datasets. Sometimes could be need necessary some manual preprocessing on data.
Black box issues	Usually proprietary tools are black-box that means you don't know how they work, sometimes is impossible understand how they manage the problem and how they achieve a result. The process of model development is hidden.
Accuracy	Depending on the problem to solve it can perform well or not. Being the model development automatic, it cannot be customized as traditional ML models are. Nevertheless, the accuracy is quite good as we will see in our experiments (Chapter 4) on a regression problem.
Reliability	AutoML is a new technology in the early stage of its life, its performances are good considering the time and effort needed but the user don't know clearly how it works and how reliable it is.

Table 7 - AutoML constraints

2 IMPACT ON DATA SCIENCE

In this chapter, the focus is on the impact that AutoML solutions have on the organization of the data science team and their professional roles. The study of impacts that new technologies have on the organizations and on the jobs is a very crucial perspective to assess because every innovation brings changes to adopt inside companies and inside the individual job of everyone. AutoML, as we will see, brings huge changes in the traditional workflow of data science and in particular in developing ML models. In particular, the analysis will discuss the workflow of ML projects considering as a reference the CRISP-DM model. Then the focus will go on the professional role of data scientist, about this role we will define its competencies and skills of today and then we will build a scenario analysis in which AutoML have different influences on the data scientist job. After the analysis of data scientists will be the turn of the analysis of the emerging role of citizen data scientist, a new professional role born with the adoption of analytics platforms like AutoML software that make easier working with data.

This is a section totally qualitative, built on the analysis of several articles published in recent years on the professional figure of data scientist and the new emerging figure of the citizen data scientist.

The chapter is structured as follow:

- CRISP-DM model with AutoML
- Deep-dive on the figure of data scientist
- New professional role: citizen data scientist

2.1 CRISP-DM model with AutoML

This model already explained in the introduction is used to define the standard steps in developing a ML project. Resuming it has 6 main steps with a circle shape since each phase is dependent on each other there are different iterations if the output of some step is not good enough. Thanks to CRISP-DM model is easy to identify the impact of AutoML on the traditional workflow of a ML project.

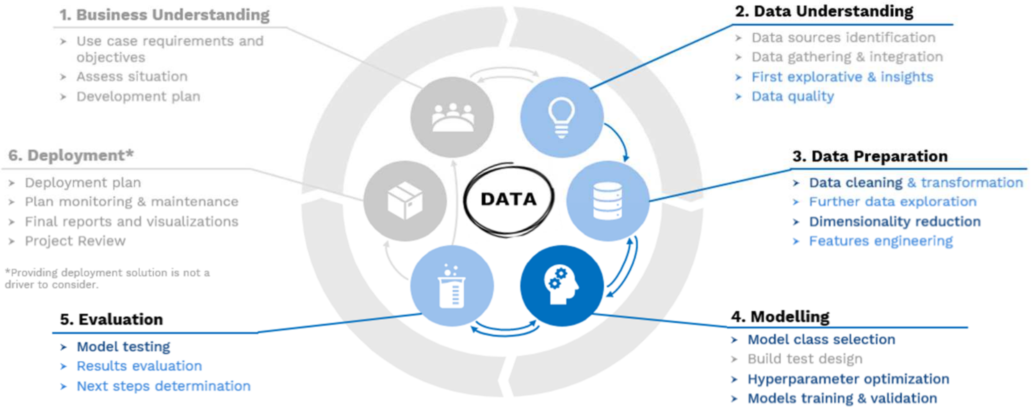

1. Business Understanding
Use case requirements and objectives
Assess situation
Development plan
2. Data Understanding
Data sources identification
Data gathering & integration
First explorative & insights
Data quality
3. Data Preparation
Data cleaning & transformation
Further data exploration
Dimensionality reduction
Features engineering
4. Modelling
Model class selection
Build test design
Hyperparameter optimization
Models training & validation
5. Evaluation
Model testing
Results evaluation
Next steps determination
6. Deployment*
Deployment plan
Plan monitoring & maintenance
Final reports and visualizations
Project Review
*Providing deployment solution is not a driver to consider.
DATA

Figure 26 – Functions automated by AutoML considering the CRISP-DM model

AutoML impacts on 4 out of 6 steps of the CRISP-DM model, leaving untouched the business understanding and the deployment phases. We didn't consider the deployment phase as a crucial driver for AutoML, and so it wasn't considered in the analysis of the CRISP-DM model. The first phase of the model, the business understanding, isn't affected because the automatic study of the problem context is not possible yet by the machine, so an expert is needed to set the problem priorities and develop a plan focused on objectives.

The classical framework is massively impacted, the whole technical part where the high-skilled competencies of a data scientist are necessary now can be automated by AutoML systems. A ML model could be developed autonomously by the machine, individuals must only set the objectives and monitor the results. The next table will explain in detail the different phases automated by AutoML:

Phase	Task	Description
Business Understanding	First exploration & insights	Usually after having uploaded the data, AutoML systems made a first explorative analysis on data, extracting useful statistics and trends.
	Data quality	AutoML systems check if the data quality is good or not to do analytics projects
Data Preparation	Data cleaning and transforming	Data are cleaned autonomously by the system based on different techniques, already seen in the (1.4.1 section)
	Further data exploration	Data are analyzed to find hidden patterns and report useful insights on their structure and relation.
	Dimensionality reduction	Large datasets are autonomously reduced by AutoML to the more representative number of features useful to design the model.
	Features engineering	AutoML automatically generate feature from those already existing, these features have the objective to better describe the behavior of some specific patterns inside data.
Modelling	Model class selection	This phase is highly automated by AutoML because it chooses the best model/algorithm to apply on your data according with the problem setting and objective.
	Hyper-parameters optimization	Once a model is selected, AutoML tries to improve its performances through the automatic combination of hyper-parameters, where hyper-parameters define the training process of the algorithm.
	Models training and validation	AutoML system generally train all the possible models on data, then the best model validated is chose based on performance metrics.
Evaluating	Model testing	The model is autonomously tested on the test set.
	Results evaluating	Each model tested is evaluated on some performance metrics, and the best fitting data never seen before is selected as the best model.

Table 8 - Phases automated by AutoML systems

2.2 Deep-dive on Data Scientist

Now that the impact on the traditional workflow is cleared, we can focus our attention on the professional role involved during the project accomplishment. Precisely this paragraph is dedicated to the data scientists, trying to answer the following demand:

- Who is he/she?
- In what consists his/her job?
- Why they are so important for companies?
- In which measure AutoML impact on them?

In 2012 the professional role of data scientist was called by Harvard Business Review:

"The Sexiest Job of the 21st Century"

This affiliation was due to the fact that nowadays we are seeing a continuous evolution of companies to become data-driven and the way they do it is dealing with data they create or find outside the company boundaries. But who will manage and extract insights from these data? The data scientist. This is why this figure is so requested by companies all over the world. The two following definition capture the essential role of the data scientist:

> *"A data scientist is an individual that performs statistical analysis, data mining and ML processes on a large amount of data to identify trends, figures and other relevant information."* **[15]**

> *"Data scientists generally analyze big data, or data depositories that are maintained throughout an organization or website's existence but are of virtually no use as far strategic or monetary benefit is concerned. Data scientists are equipped with statistical models and analyze past and current data from such data stores to derive recommendations and suggestions for optimal business decision making."* **[15]**

It is important to highlight that not all the data scientists are equal since the data scientist skill-map is very wide, someone could be very wise on statistics and others in computer science or even in business. What is very important is having different data scientists with different knowledge profiles in order to have a team prepared for many different types of problems. Since there are many roles with different backgrounds in data science, it is important to build heterogeneous data science team with different skills. The following table list the 6 main roles involved in a Data Science Team highlighting which differences there are between the different roles and which competencies are requested to define a specific role.

The traditional framework of a team is as follow:

	Data Scientist	***Data Analyst***	***Process Expert***	***Data Architect***	***Data Engineer***	***Data Science Manager***
Role	Clean and organize data, create ML models, actively supports presentations or results	Collect, process and perform statistical data analysis	Improve business processes as intermediary between business and IT	Create blueprints to integrate data mgmt systems; Centralize, protect and maintain data sources	Develop, construct, test and maintain architectures	Manage a team of data analysts, data engineers and data scientists
Languages and tools	R, Python, SAS, Matlab, KNIME, SPSS, Tableau, SQL, Hive, Pig, Spark	SQL, KNIME, SPSS	SQL, Tableau	SQL, XML, Hive, Pig, Spark	SQL, Kive, Pig, R, SAS, Python, Java, Ruby, Perl	SQl, R, SAS, Python, Matlab, SPSS, KNIME
Skills and talents	Distributed computing, predictive modelling, cognitive computing, storytelling & visualizing, math, stats, ML and data mining	Spreadsheet tools, D systems, communication & visualization, math, stats	Basic tools, data visualization tools, conscious listening & storytelling, BI understanding, Data modelling	DWH solutions, Deep knowledge of DB architecture ETL, spreadsheet & BI tools Data modelling System development	DB systems, Data modelling & ETL tools, Data APIs, DWH solutions	DB systems, Leadership & PM Interpersonal communication, Data mining & predictive modelling

Table 9 - Data Science professional roles

It is obvious that data science professional figures have strong competencies in different fields and the team composition is structured to cover all the skill needed to deliver high-value projects. Each role has its own objectives and area of competence. Usually, each role is the actor in a specific part of the pipeline of ML.

With the recent arise of AutoML there are many discussions about the possible future scenarios that this innovation could bring inside data science organization. During the researches we noticed that there are two main paths:

1. AutoML is considered as a technology in support to data scientists
2. AutoML is considered as a technology that is going to substitute data scientist work

With a scenario analysis we are going to see in detail what are the threats and opportunities that AutoML is bringing to the world of data science in both cases.

1) AutoML: a support tool for data scientists

The first scenario is one in which AutoML is considered as a support tool for data scientists in doing their work. AutoML enables data scientists to avoid repetitive and no value-added tasks. Data scientists can focus just on high-value functions where their expertise is really needed. Considering that the 80% of their time is used to make data preparation (Chapter 1.5) we can understand that using AutoML to automate the tedious tasks let data scientists to increase their productivity because all the time that was used to clean data and to engineer data, now it can be used to create and to train models.

Furthermore, AutoML can be used as a starting point on which develop very accurate models. In this perspective data scientists run AutoML to obtain the best starting model and then they can improve the model tuning hyper-parameters or selecting different features to consider. It depends on which tool is used and their functionalities, if we are taking in consideration a tool able to automate all the pipeline it can return even the final model but if we consider a 'narrow' tool it can deal only with specific tasks of the pipeline.

In this scenario, the main benefits are:

- Increase the productivity of data scientists
- Decrease time to develop a model
- Good starting point (initial model to improve)
- Avoid human biases

The main constraints are:

- Some tools are black-box and data scientists don't know how they create models
- Difficulty for AutoML to understand the domain of the problem, today there is a need for human supervision mainly for problem setting.

2) AutoML: a threat for data scientists

This scenario is the worst possible for data scientists. This scenario assumes that AutoML in the next future will substitute data scientists' job. This is a scenario in which ML is totally automated by AutoML solutions without the need of data scientists inside companies that rely on the automated systems their decisions. The crucial factor behind this possible scenario is the reliability of AutoML systems, if these systems will achieve performances as good as data science teams in developing ML models, the probability that many companies will adopt these solutions instead of hiring data scientists is quite high because the cost of a team of data scientists is important and even the time to create a model is to consider as an important cost. Instead, an AutoML solution for a company is less expensive in terms of money, time and human resources. In this scenario, if the different AutoML solutions will become the standard used in ML projects, the very high skilled and costly figure of data scientists will be irrelevant if the only task to do is to define the objective and upload the dataset on the software.

Of course, this will be not an immediate revolution, but it will be a constant evolution of roles and competencies. The figure of data scientist can become replaced from a technical point of view by AutoML systems and from an organizational point of view by the new professional figure of citizen data scientist. This emerging professional role will be analyzed in the following section (Chapter 2.3).

The benefits of this scenario are:

- Companies will save money, time and human resources
- Arising of new professional figures, creation of new job
- Automation ensures no human biases
- Democratization of ML

The constraints are:

- Data scientists will lose importance
- Knowledge inside systems and not inside humans' mind
- Competition problem, the same tool can be used by many companies in the same industry

These are the two main flows of thought about the future of data science and in ML world. These two scenarios were defined by the analysis of the following articles:

- 3 Reasons Why AutoML Won't Replace Data Scientists Yet – KDNuggets **[16]**
- Automated machine learning: just how much? – KDNuggets **[17]**
- Does AutoML work for all data science stakeholder: expectations vs reality – AIM **[18]**
- Is AutoML the Answer to the Data Science Skills Shortage? - InformationWeek **[19]**
- AutoML Tools Emerge as Data Science Difference Makers – datanami **[20]**
- The Risks of AutoML and How to Avoid Them – Harvard Business Review **[21]**
- Why data-scientists are rejecting automation (AutoML) – Kortical **[22]**
- Implementing Automated Machine Learning (AutoML) – Forbes **[23]**
- Does Google AutoML eliminate the need for ML specialists? – Quora **[24]**

From the analysis of many articles we can say that the general idea is that thanks to AutoML tools that allow data scientists to avoid repetitive and time-consuming tasks along the ML pipeline, this figure will change the focus of its tasks a lot in the next future. We can suppose that data scientists will maintain all its competencies since that every day there are new problems to face and the expertise in computer science, mathematics and statistics are fundamental to solve future challenges, but mainly to define the problem set.

What will change for '*today*' data scientist will be the focus of their work. Indeed, tedious and repetitive tasks will be automated by AutoML, leaving the data scientist all the time to focus his attention on complex problems that need his deep expertise. For example, data scientists could save time automating the data cleaning and spend all the efforts in the modeling phase, choosing algorithms and tuning hyper-parameters to achieve outperforming results.

The most probable solution about the future is the intersection of the two possible scenarios, where AutoML systems will become an important support tool for the job of data scientists increasing their productivity and at the same time AutoML will enable the arising of the new professional roles like the citizen data scientist to fill the gap between the demand for data professionals and their availability.

2.3 New professional role: Citizen Data Scientist

"Citizen Data Scientist," a term coined by Gartner, refers to advanced data analytics professionals or data professional that needs or wants to implement ML technology. A citizen data scientist is a role that analyzes, creates data and business models for their companies with the help of Analytics systems and technologies. Citizen data scientists do not necessarily need to be data science or business intelligence experts. This role is given to employees in an organization who can use the analytics tools and technology to create ML models.

The role of citizen data scientists was created as companies faced shortages of trained data scientists. While this new role is not a substitute for data scientists, it has proved effective ability in fulfilling the purpose for which it was created. New tools and technologies are being rolled out to fill the void created by the scarcity of data scientists. Such tools could create data models and provide deep insights as well, so companies have trained people to handle these tools. While citizen data scientists are not considered experts in data science, they are able of using the tools to provide various insights that can be useful for businesses in which they are actors.

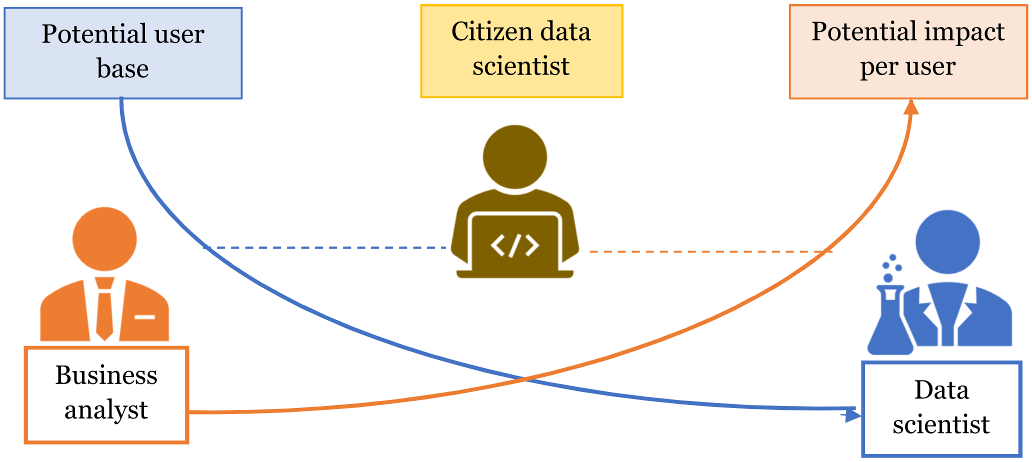


Figure 27 - Citizen data scientist profile

Figure 27 easily explains the positioning of the citizen data scientist, it is in the middle between two well defined professional figures: the business analyst and the data scientist. The citizen data scientist has some skills more than the business analyst but not as much as the data scientist, and the user base is larger than the data scientist but lower than the business analyst. Citizen data scientists are not intended to replace data scientists. In fact, both roles can work in tandem. While data scientists can research and find novel ways of creating data insights, citizen data scientists can continue to use the analytical support systems like AutoML. As we can see from the image a citizen data scientist is an intermediary role between the business analyst and the data scientists.

About this new arising figure were written different articles, we select the most reliable ones from three top companies: Gartner, Forbes, and Forrester. These three articles try to define the edges of this new role, for each article here below there is a summary of the concepts treated. *(In the webography, pg. 123, there is the list with all the papers analyzed to define this new emerging professional figure)*

Gartner (May 13, 2018 - by Carlie Idoine) **[25]**

Gartner defines a citizen data scientist in "citizen data science augments data discovery and simplifies data science" as a person who creates or generates models that use advanced diagnostic analytics or predictive and prescriptive capabilities, but whose primary job function is outside the field of statistics and analytics.

Citizens data scientists are "power users" who can perform both simple and moderately sophisticate analytical tasks that would previously have required more expertise. Today, citizen data scientists provide a complementary role to expert data scientists. They do not replace the experts, as they do not have the specific, advanced data science expertise to do so.

Gartner identifies 7 main skills related to the citizen data scientist:

1. A contextualized vision of the organization
2. Unique perspective of individual business area
3. Proven applicability of analytical techniques to business problems
4. Able to go to bat to justify business value
5. Appetite for what matters relative to business priorities
6. Been around the block and has connections
7. Involved hands-on in multiple analytic areas and activities

Many forces are contributing to feeding the potentially disruptive and transformative power of this emerging citizen data scientist role:

- Organizations are increasingly prioritizing the move into more advanced predictive and prescriptive analytics.
- The expert skills of traditional data scientists to address these challenges are often expensive and difficult to come by. Citizen data scientists can be an effective way to mitigate this current skills gap.
- Technology is a key enabler of the rise of the citizen data scientist now. Technology has gotten easier for non-specialists to use. Analytics and Business Intelligence tools are extending their reach to incorporate easier accessibility to both data and analytics.
- Technology development also includes augmented analytics, often referred to as AutoML tools.

These four last points explain the enhancing need of citizen data scientist figures in companies. This article sustains the scenario in which AutoML enable even a non-data scientist to perform analytical tasks, but data scientists keep their importance unvaried. AutoML is the enabler of a new professional figure.

Forbes (Feb 20, 2019 – by Marc Zionts) **[26]**

This article talks about the gap between the demand of experts in data science and their availability. Today every company, in order to survive and beat competitors collects, processes and analyzes data, but to do these tasks there is a need of the professional figure of data scientists. These figures are in charge of data wrangling, discovery, analysis, structuring, cleaning, validating and communicating data for projects and company needs.

The problem is that the available number of data scientists cannot satisfy the need of the market for these figures.

In order to cover this gap inside companies, people must deal with data even with many different roles, thanks to software tools and new technologies today even employees without strong knowledge in data science can deal with deal. This fact helps bridge the supply and demand gap of expert data scientists.

Forbes identifies business professionals using analytical software as Citizen Data Scientists.

With the increase of data interaction, it's imperative to set citizen data scientists up for success when it comes to understanding, communicating and acting on data. They have some data science skills but are not as advanced as data scientists. They are 'power-user' that can perform both simple and moderately sophisticated tasks on data, always thanks to software that sometimes do not need any line of code to be used.

The added value of this new professional role is that they have a unique perspective in their specific business area, simplifying the problem setting phase and defining what are the results to obtain with a very focused view.

Any data-driven business, regardless of the size, demands resources for extracting and acting on meaningful insights to further the position of their company.

Even from this second article, the figure of citizen data scientist is treated as necessary for the future of the companies, enabling these latter to become data-driven. Citizen data scientists will fill the demand for lack of data scientists.

Forrester (Feb 8, 2019 – by Mike Gualtieri) **[27]**

Everyone is talking about the Citizen Data Scientist, but no one can define it.
The simplest definition given by Forrester of a citizen data scientist is:

“Non-data scientist.”

It’s not a pejorative, it just means that citizen data scientists nobly desire to do data science but are not formally schooled in all the ins and outs of the data science life cycle. Forrester makes an example useful to understand the difference between a data scientist and a citizen data scientist:

A citizen data scientist may be quite savvy about what enterprise data is likely to be important to create a model nut may not know the difference between GBM, random forest, and SVM (Support Vector Machine). Those algorithms are data scientists’ geek-speak to many of them. The citizen data scientist’s job is not data science; rather, they use it as a tool to get their job done.

Then Forrester enlarge the definition of citizen data scientist saying that it is:

“A business person who aspires to use data science techniques
such as ML to discover new insights and create predictive models
to improve business outcomes.”

According to Forrester, they must learn the ML lifecycle: data acquisition, data preparation, feature engineering, algorithm selection, model training, model evaluation and finally insights and/or predictions.

They even must learn to program in R or Python. If they are lucky, they will download RapidMiner, KNIME, tools that provide nice visual drag-and-drop interfaces versus harsh coding.

From this last article, the citizen data scientist is described majorly under the technical point of view, he/she are not data scientists but with the help of AutoML and other analytical systems can deal with data. In order to become a citizen data scientist is necessary knowing at least the machine learning life cycle and then they must develop and increase their knowledge in some programming language like Python or R.

All the articles taken in consideration have talked about citizen data scientist, not as a threat for data scientist, instead it is considered to become a very important figure for companies because it will fill the gap between the demand of data science professional roles and the actual availability, the citizen data scientist is the figure in charge to help organizations to become strategically data-driven. AutoML will give the power to them to deal with data and develop a meaningful ML model to make prediction and classification tasks.

3 AUTOML MARKET ANALYSIS, INNOVATION AND BENCHMARKING

In this chapter, we will analyze what are the characteristics of AutoML systems, starting with a description of the market and then with a classification about the main AutoML providers. Before talking about the “players” we will talk about AutoML from a theoretical point of view, explaining why it can be considered innovation. In the theoretical part we are going to apply two famous models to the topic of AutoML:

- Definition of the AutoML: Technology-Push vs Market-Pull
- Red Ocean vs Blue Ocean, innovation strategy of AutoML

After having identified AutoML technology from an academic perspective with the two models selected, we will analyze the different systems, trying to show the actual situation in the global market and classifying them to create a comprehensive big picture that contains all the main available AutoML tools in existence today. The Big picture is a comprehensive table in which are listed all the main AutoML solutions available in the market worldwide.

Then the focus will go on the analysis of tools’ customer bases, with the purpose to define a framework that shows which kind of companies need this service, and in the end, we will build a benchmark framework to classify deeply AutoML systems, then applied on the five AutoML tools tested in the business case in Chapter 4. These tools used for building the benchmarking are the core of the final part of the thesis, the business case, a real ML success case developed in Bip, then simulated by AutoML with the aim to highlight the differences between ML project and AutoML projects, the tools considered are Google Cloud AutoMl Tables, Dataiku, Azure, AWS Sagemaker and H2O Driverless AI.

3.1 AutoML: Innovation & Strategy

Let's define innovation:

> *"To be called an innovation, an idea must be replicable at an economical cost and must satisfy a specific need. Innovation involves deliberate application of information, imagination and initiative in deriving greater or different values from resources, and it includes all processes by which new ideas are generated and converted into useful products. In business, innovation often results when ideas are applied by the company in order to further satisfy the needs and expectations of the customers."* **[28]**

Let's define strategy:

> *"Strategy is a high-level plan to achieve one or more goals under conditions of uncertainty. In the sense of the "art of the general," which included several subsets of skills including tactics, siegecraft, logistics etc."* **[29]**

> *"The art and science of planning and marshalling resources for their most efficient and effective use."* **[30]**

From the definitions of both innovation and strategy is easy to understand that AutoML perfectly fits with these terms because it is a product/service that has a commercial value and a business model behind with the aim of satisfying different needs. AutoML brings innovation because of breaks the traditional ML approach, both for companies and for data scientists.

The reason why it is an innovation and not an invention is that today it has been created with a commercial value, before the development of these systems the idea of automating all the processes and let the machines to do everything already existed. The innovation is that today everyone can access these resources through different software.

Red Ocean vs Blue Ocean

It is interesting studying the AutoML phenomenon considering the theory of Red Ocean and Blue Ocean. Being a digital innovation in the field of data science and being considered as a product/service with a growing market, it can be described under the strategic point of view for developers. Before trying to define the AutoML belongness to one or the other "ocean", let's define the theory behind this model. (W. C. Kim, R. Mauborgne, 2004)

Red Ocean

With red ocean, we mean all the companies in existence today, the known market space.

> *"The key goals of the red ocean strategy are to beat the competition and exploit existing demand."*

A company belongs to this group if its strategy relies just on beat the competition, trying to outperform the rivalries. Companies inside the red ocean do not innovate the business model, they fight to grab market share to others. Typically, the strategy is on price. A red ocean is already full of players, it is difficult to enter and win the market for newcomers.

Blue Ocean

A blue ocean strategy is based on creating demand that today doesn't exist, rather than fighting over it with other companies. In order to understand the potentiality of this strategy, you must keep in mind that there is a deeper potential of the marketplace that hasn't been explored yet. Generally, blue oceans are created inside red ocean by expanding existing industry boundaries.

> *"The key goals of the blue ocean strategy are finding the right marketing opportunity and making the competition irrelevant."*

The main differences between the two strategies are the following:

Red Ocean Strategy	Blue Ocean Strategy
Compete in existing market	Create uncontested markets to serve
Beat the competition	Make the competition irrelevant
Exploit existing demand	Create and capture new demand
Make the value-cost trade-off	Break the value-cost trade off
Align the whole system of a firm's activities with its strategic choice of differentiation or low cost	Align the whole system of a firm's activities in pursuit of differentiation and low cost

Table 10 - Differences between a blue ocean and a red ocean strategy

According to the main characteristics associated to both strategies, it is evident that AutoML perfectly fits well with the features regarding the blue ocean strategy. The reasons for this affiliation are listed below:

- AutoML created a new uncontested market, a market in which the demand for those systems that allow everyone to deal with ML is growing even more. This new service created a new uncontested market space.
- Creating a new market in which companies without knowledge in ML now can develop ML activities and the fact that there are few AutoML providers worldwide developing different systems with different purposes makes the competition irrelevant.
- AutoML has the potential to create a huge demand because today every company is going to become data-driven and thanks to AutoML the change can be faster

and less expensive under the perspective of human resources and internal knowledge.

- AutoML breaks the value-cost trade-off because AutoML services/tools can be purchased at low prices, providers aim to achieve both differentiation and low price for customers. In this way, the demand is increasing because the wide range of solutions attract more customers and the low price is a strong attractive factor too.

Google is providing a wide range of AutoML solutions **[31]**:

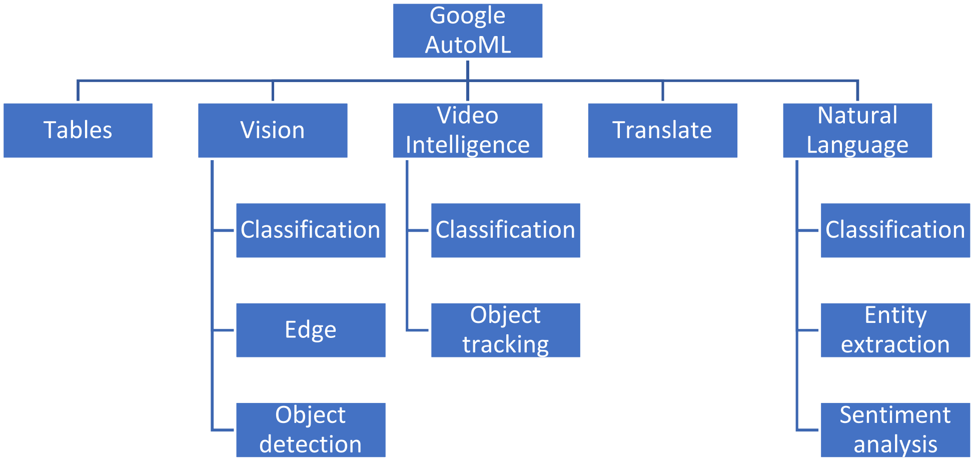


Figure 28 – Google cloud AutoML service diversification

As we can see Google is focusing on diversification of offer keeping low prices, even because today its AutoML services are still in Beta.

Concluding we can state that AutoML can be classified as a red ocean strategy because it has enlarged the boundaries of an existing industry creating new demand for an innovative service/product.

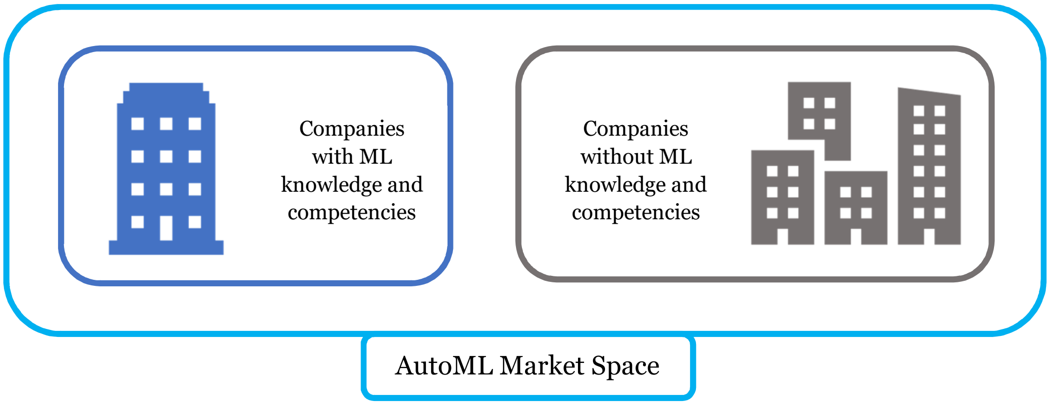

Companies with ML knowledge and competencies
Companies without ML knowledge and competencies
AutoML Market Space

Figure 29 - Blue Ocean strategy of AutoML

The picture is the representation of the democratization of ML practices in the market. It shows the Blue ocean effects that create a link between companies with ML knowledge and companies without those competencies, creating a new market space.

3.2 AutoML: Market Pull

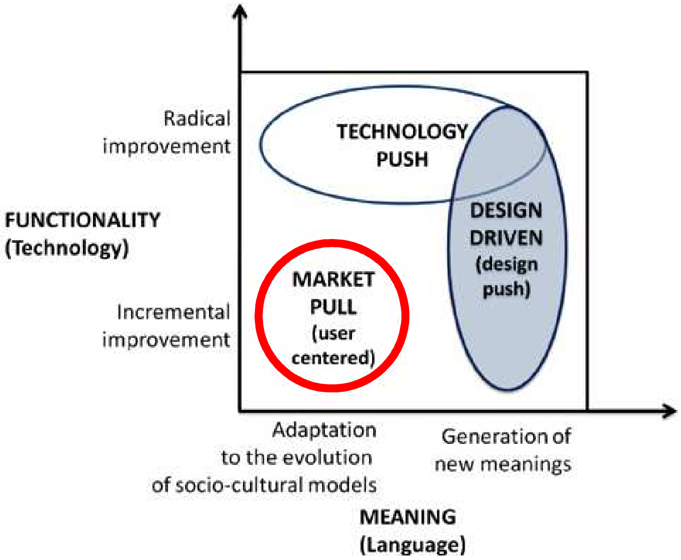

Radical improvement
FUNCTIONALITY (Technology)
Incremental improvement
TECHNOLOGY PUSH
DESIGN DRIVEN (design push)
MARKET PULL (user centered)
Adaptation to the evolution of socio-cultural models
Generation of new meanings
MEANING (Language)

Figure 30 - Innovation model to define a new technology

AutoML is a solution born to solve different needs present in the market of data science. AutoML is taking more and more importance due to the benefits it brings. It can be classified as a market pull strategy because it is a need not created by the product but created by problems dealing with ML. Market pull means that the new product/service/innovation is led by the demand of the market, the market has some needs that must be solved and so there are all the conditions to develop new solutions to fill the gaps inside the market. In our case, AutoML tools try to fill different gaps and problems of the traditional ML. AutoML is not a radical innovation because it uses all the methodologies already known in 'ML' but it adds the high-value concept of automating the entire pipeline, it aims to let the machines understanding how to autonomously learn patterns from data and choose the best model that describe the patterns for data in consideration. It is continuous innovation in the field of ML because in the last years there wasn't a unique solution that solved all the problems, till now there are continuous investments in this field confirmed by the fact that companies like Google had launched in Beta different services of AutoML.

The problems in the market that AutoML tries to solve are:

- To increase the productivity of data scientist work by automating the repetitive and not added value activities.
- To leverage big data to drive business, improving the data-driven approach of companies.
- Democratization of ML, enabling even companies without ML competencies to leverage their business on ML model to extract data and insights to run their businesses.

It is focused on the user because it can be considered a support tool to use in ML project, especially, it is helpful for data scientists in some part of the pipeline to perform tedious and repetitive tasks but also very useful for non-data scientists that with a simple system can develop ML model without any line of code. The user is the beneficiary. Indeed, many systems have user-friendly interface, usually drag and drop, to permits any user to understand and to execute difficult analysis. It is helped from the trend of companies of becoming data-driven because it could help them in extracting value from their data without having strong competencies inside. The need of automating the ML is caused by the complexity of the ML itself, it needs a lot of time and competencies to achieve good results, and those results are very important for companies that gain many advantages against their competitors.

3.3 AutoML Market Research: Methodology

To build the big picture of the market, the analysis was performed on all the AutoML providers that nowadays are selling their products, for which there is a sufficient amount of available documentation.

The market analysis has been developed in parallel from two different point of views:

- Classification of the tools based on its nature.
- Classification of the tools based on its capabilities.

Classifying a tool on its nature means to understand if it was developed by a company with a commercial purpose or if it was developed with research purposes. This type of classification is the first that we are going to explain. The second classification, based on the capabilities of each tool, aims to build a framework able to highlight what kind of problems each tool can solve.

As we will see, the two classifications are correlated and joining them we will create a comprehensive table in which each tool will be classified based on both the classifications.

To define the nature of the tool of AutoML was developed a simple but useful framework. Two Boolean variables that create four categories, the type of service could be 'Narrow' or 'Generalized' and the type of software could be 'Open Source' or 'Proprietary'. Combining in a matrix of these two variables we have created 4 categories to classify the AutoML tools.

Table 11 - AutoML tools classification framework

		Type of software	
		Open Source	***Proprietary***
Type of service	***Narrow***	Narrow Open Source	Narrow Proprietary
	Generalized	Generalized Open Source	Generalized Proprietary

To give a reason to this classification, we are going to explain better the concept between type of service and type of software.

Type of service: Narrow service

With the term ‘narrow’ we mean to identify all the tools that are not delivering a service of automating the whole ML pipeline but just one or more parts of it. Having subdivided the pipeline into four main phases (feature cleaning, feature engineering, model selection, and hyper-parameters optimization), we consider ‘narrow’ a tool that automates at least one up to three phases of the pipeline.

Type of service: Generalized service

Classifying a tool as Generalized means that it executes all the ML pipeline, all the work is made by the machine and no code is needed in any part. Usually, these tools ingest the various datasets, you set the problem to solve and the tools do all the process for you, delivering to you the best model is calculated. We consider generalized a tool when it executes all the four steps we fixed.

Type of software: Open source

The term refers to something accessible to everyone, publicly shareable and available. In the context of software, “open source” development stands for a specific approach to create computer programs. Open-source software is software with source code that anyone can inspect, modify, and enhance. The objective of the open-source choice is to evolve the code thanks programmers and developers that can put their hands on. It has the main purpose to take ahead the research field without commercial aims. For the open-source solutions were taken into account just those one consolidated, with good documentation and with use cases mainly found on GitHub.

Type of software: Proprietary

Proprietary software is any software copyrighted and bears limits against use, distribution, and modification. Proprietary software is primarily commercial software that can be bought, leased or licensed from its vendor/developer. It can be purchased the license for a fee, but relicensing, distribution or copying is prohibited. For proprietary tools were toke into account those that are already in the market, they are already an alternative for customers with a defined offer. These are the main tool used for AutoML activities, many of them are proprietary but there is a good portion of open-source. This means that AutoML is seen as a business opportunity by companies that invest and develop tools to sell and at the same time there is a high interest in developing and improving the actual AutoML solutions since there are a lot of people that through the open source software give his/her contribute.

The number of AutoML systems considered in the research is 53. It can be insightful to understand the development trend, with a trend we mean the class of belongingness of each tool created.

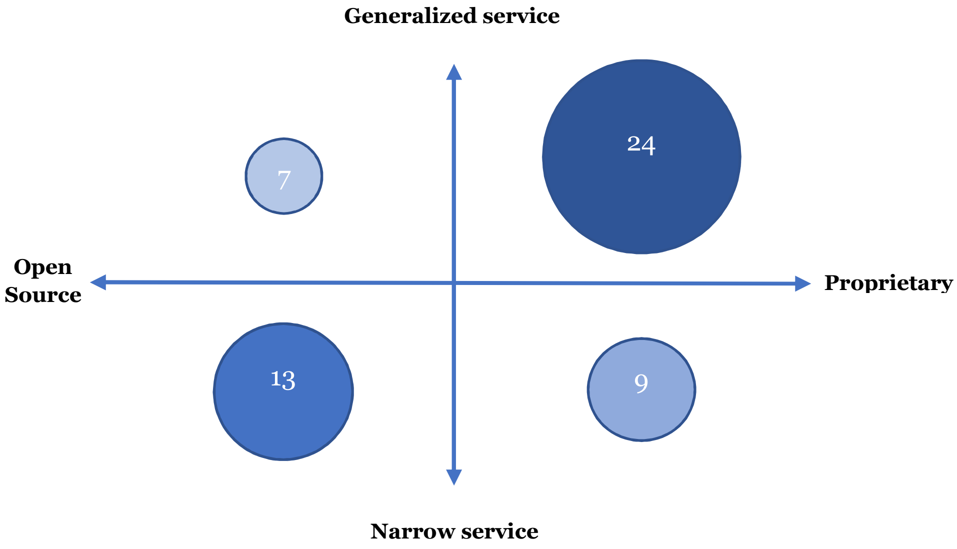


Figure 31 - AutoML classification map

To have a whole picture of the development trend we can use this positioning map to highlight the correlation between the proprietary tools offering generalized service and the correlation between open source tools offering narrow service. From this first analysis emerged that proprietary tools aim to automate all the pipeline to be more valuable by customers (total of 24 tools), instead open source solutions aim to create knowledge on specific tasks of the pipeline and create documentation for the crowd and from the crowd (total of 20 tools). The other two segment are in considerable minority, indeed there are only 7 open source tools that automate the entire pipeline and only 9 proprietary tools that are narrow.

With the positioning map we conclude the first classification based on the nature of the tools, now it is time to introduce each tool with their functionalities.

The following of the chapter will contain:

- A table with all AutoML tools considered. The tools will be characterized by the activity they automate in the ML pipeline and their classification made in this section.
- AutoML customers analysis, considering an amount of 229 confirmed customers.
- A deep analysis about AutoML systems tested in the business case.

3.4 Market Big Picture

The analysis of the different AutoML tools was conducted considering the ML phase that the tools can automate, so to set a consistent framework to the analysis, the pipeline was divided in 4 main phases.

With this structure is easy to understand what kind of functionalities a tool can perform:

- Data cleaning
- Features engineering
- Hyper-parameters optimization
- Model selection

The pipeline was split in these 4 phases to group different techniques performed on data by different tool, it is a simplification used to classify tools with same functionalities even if the technical processes can differ from one to another.

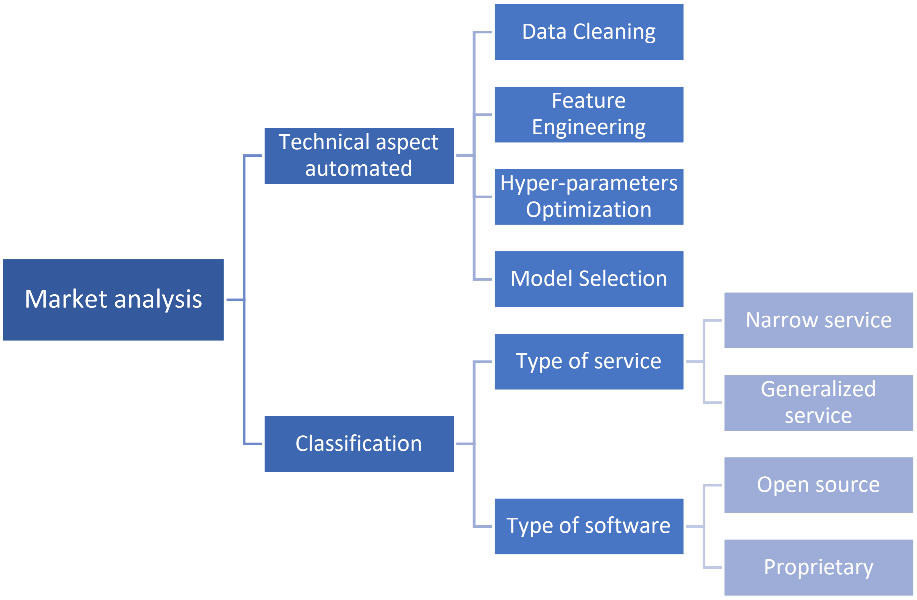


Figure 32 - Market analysis framework

These four phases were treated already in Chapter 1.4 when we described the pipeline of a ML project. In this way we have a clear vision about the capabilities of each tool.

This part of the thesis was developed making one assumption: automated ML is a concrete product/service to sell and not an extension of ML services, it is not considered as a complementary product/service. It is considered not as an extension of ML services but like a stand-alone product/service.

First, the following table give a comprehensive overview of the main tools considered in the analysis with both their functionalities and classifications. The table is sort first by open source and then by proprietary tools in alphabetic order:

AutoML Tools	Data Clean.	Feat. Eng.	Hyper-params Opt	Model Select.	Class	Year	Reference (Link)
ATM – Auto Tune Models (MIT)			✓	✓	Narrow open source	2019	https://github.com/HDI-Project/ATM
Auto-Keras			✓	✓	Narrow open source	2018	https://autokeras.com/
Auto-sklearn	✓	✓	✓	✓	Generalized open source	2018	https://github.com/automl/auto-sklearn
Boruta-py		✓			Narrow open source	\	https://github.com/scikit-learn-contrib/boruta_py
Categorical-encoding		✓			Narrow open source	\	https://github.com/scikit-learn-contrib/categorical-encoding
ENAS-Pytorch			✓		Narrow open source	\	https://github.com/carpedm20/ENAS-pytorch
FeatureHub	✓	✓			Narrow open source	\	https://github.com/HDI-Project/FeatureHub
Featuretools		✓			Narrow open source	2017	https://www.featuretools.com/
H2O automl			✓	✓	Narrow open source	2012	http://docs.h2o.ai/h2o-tutorials/latest-stable/h2o-world-2017/automl/index.html
HpBandSter			✓		Narrow open source	2018	https://github.com/automl/HpBandSter
Hyperopt			✓		Narrow open source	2018	https://github.com/hyperopt/hyperopt
MLBox	✓	✓	✓	✓	Generalized open source	2017	https://github.com/AxeldeRomblay/MLBox
Pybrain			✓	✓	Narrow open source	2015	http://pybrain.org/
RECIPE	✓	✓	✓	✓	Generalized open source	2017	https://github.com/laic-ufmg/Recipe
Tsfresh		✓			Narrow open source	\	https://github.com/blue-yonder/tsfresh
TPOT	✓	✓	✓	✓	Generalized open source	2015	https://github.com/EpistasisLab/tpot
Trane		✓	✓	✓	Narrow open source	\	https://github.com/HDI-Project/Trane
TransmogrifAI salesforce	✓	✓	✓	✓	Generalized open source	2018	https://github.com/salesforce/TransmogrifAI
Aible	✓	✓	✓	✓	Generalized proprietary	2018	https://www.aible.com/blueprints/
Alteryx			✓	✓	Narrow proprietary	2011	https://www.alteryx.com/products/alteryx-platform/alteryx-designer
Auger.ai	✓	✓	✓	✓	Generalized proprietary	2016	https://auger.ai
AWS Sagemaker			✓		Narrow proprietary	2017	https://aws.amazon.com/it/blogs/aws/sagemaker-automatic-model-tuning/
Azure ML Microsoft	✓	✓	✓	✓	Generalized proprietary	2014	https://docs.microsoft.com/it-it/azure/machine-learning/service/concept-automated-ml
Big ML OptiML			✓	✓	Narrow proprietary	2018	https://bigml.com/releases/winter-2018
Big Squid	✓	✓	✓	✓	Generalized proprietary	2018	https://www.bigsquid.com/kraken/analyze
Clarifai	✓	✓	✓	✓	Generalized proprietary	2013	https://www.clarifai.com/predict
Compellon	✓	✓	✓	✓	Generalized proprietary	2010	https://www.compellon.com/technology/

Table 12 - Table listing all the AutoML tools considered in the market analysis (continue in next page)

DarwinAI	✓	✓	✓	✓	Generalized proprietary	2017	https://darwinai.ca/features.html
Dataiku	✓	✓	✓	✓	Generalized proprietary	2013	https://www.dataiku.com/dss/index.html
DataRobot	✓	✓	✓	✓	Generalized proprietary	2012	https://www.datarobot.com/
DMway	✓	✓	✓	✓	Generalized proprietary	2013	http://dmway.com/
dotData	✓	✓	✓	✓	Generalized proprietary	2018	https://dotdata.com/
Firefly.ai	✓	✓	✓	✓	Generalized proprietary	2016	https://firefly.ai/
Google Cloud AutoML	✓	✓		✓	Generalized proprietary	2019	https://cloud.google.com/automl-tables/docs/
H2O Driverless AI	✓	✓	✓	✓	Generalized open source	2017	https://www.h2o.ai/products/h2o-driverless-ai/#features
KNIME Analytics Platform	✓	✓	✓	✓	Generalized open source	2006	https://www.knime.com/blog/how-to-automate-machine-learning
Kogentix AMP	✓	✓	✓	✓	Generalized proprietary	2015	https://www.cloudera.com/solutions/gallery/kogentix-automated-machine-learning-platform-amp.html
MLJAR			✓	✓	Narrow proprietary	2016	https://mljar.com/
Neuton (Bell Integrator)			✓	✓	Narrow proprietary	2018	http://bellintegrator.com/Neuton
OneClick.ai	✓	✓	✓	✓	Generalized proprietary	2017	https://www.oneclick.ai/product/
Predictive Layer			✓	✓	Narrow proprietary	2014	https://www.predictivelayer.com/
Purepredictive	✓	✓	✓	✓	Generalized proprietary	2011	https://www.purepredictive.com/
R2.ai	✓	✓	✓	✓	Generalized proprietary	2015	https://r2.ai/product
RapidMiner	✓	✓	✓	✓	Generalized proprietary	2018	https://rapidminer.com/
SAP Predictive Analytics	✓		✓	✓	Narrow proprietary	2015	https://www.sap.com/italy/documents/2015/05/280754e0-247c-0010-82c7-eda71af511fa.html
SAS				✓	Generalized proprietary	2017	https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html#filterlist=contenttypedocwhite-paper
SigOpt			✓	✓	Narrow proprietary	2014	https://sigopt.com/product/by-model-type/machine-learning/
Sparkcognition Darwin	✓	✓	✓	✓	Generalized proprietary	2018	https://www.sparkcognition.com/product/darwin/
Squark Seer			✓	✓	Narrow proprietary	2017	https://squarkai.com/squark-seer/
Tazi.ai	✓	✓	✓	✓	Generalized proprietary	2015	https://www.tazi.ai/technology/
TIBCo Data science	✓	✓	✓	✓	Generalized proprietary	2019	https://www.tibco.com/products/data-science
Watson ML IBM	✓	✓	✓	✓	Generalized proprietary	2017	https://www.ibm.com/cloud/watsonstudio/autoai?mhsrc=ibmsearch_a&mhq=automl
Xpanse AI	✓	✓	✓	✓	Generalized proprietary	2015	https://xpanse.ai/

Table 13 - Table listing all the AutoML tools considered in the market analysis

From the table there is an evidence trend for AutoML tools to deal with hyper-parameters optimization and model selection, instead the first and the second activities are less embedded in this system because are tasks in which the context is very crucial, and a human supervision is quite important to avoid mistakes in the problem setting. Of course, many tools try to execute data cleaning and feature engineering phases, but them are critical tasks where the tools have some difficulties in understanding the context.

As we can see from the table usually open source tools tends to perform and automate just one phase of the ML pipeline, instead proprietary tools are trying to deal with the entire workflow usual for data scientists during a ML project mainly because the more they automate the pipeline the more they are valuable for clients who need a tool that perform ML tasks even without human expertise in data science and ML. Having collected the released year for each tool (for some open source tools are not available the release date) we can create a timeline in which we highlight the birth of this systems:

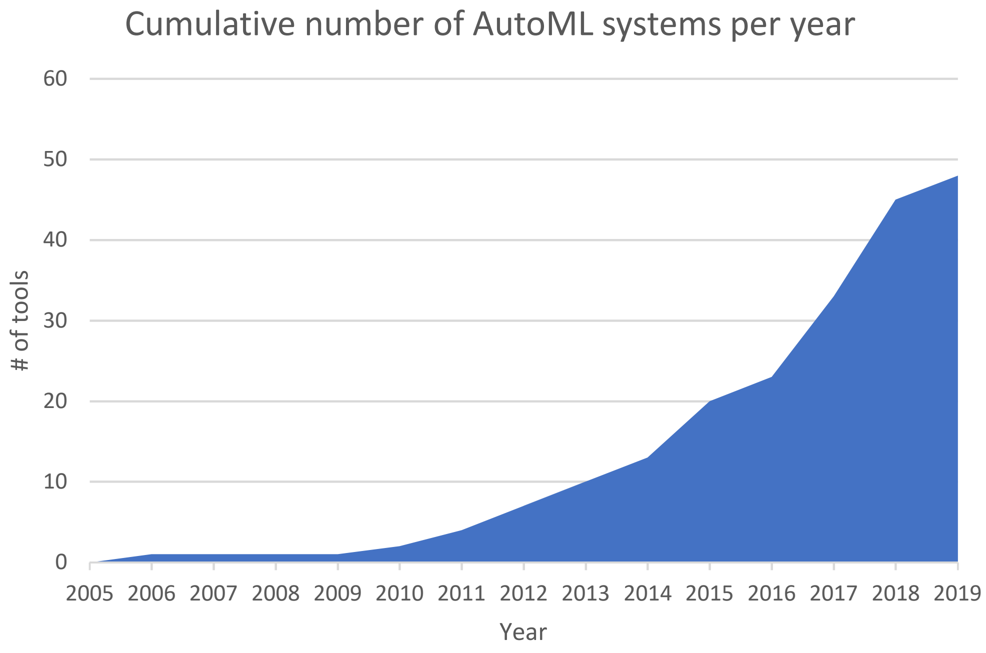


Figure 33 - Cumulate of AutoML tools products per year

In the last 5 year we have seen an increment of AutoML tools of +269%, having in 2014 an amount of 13 tools and in 2019 a total of 48 different solutions. It is a clear signal of growth regarding these new systems.

Another important result from the analysis of AutoML systems is the geographical collocation in which these systems were created. Considering our 53 tools, 20 are open source and thus they can't be associate to a geographical area of development, instead for the other 33 proprietary tools was possible define where they arose around the world. To define the geographical are of development was considered the headquarter of the companies. What emerged is that:

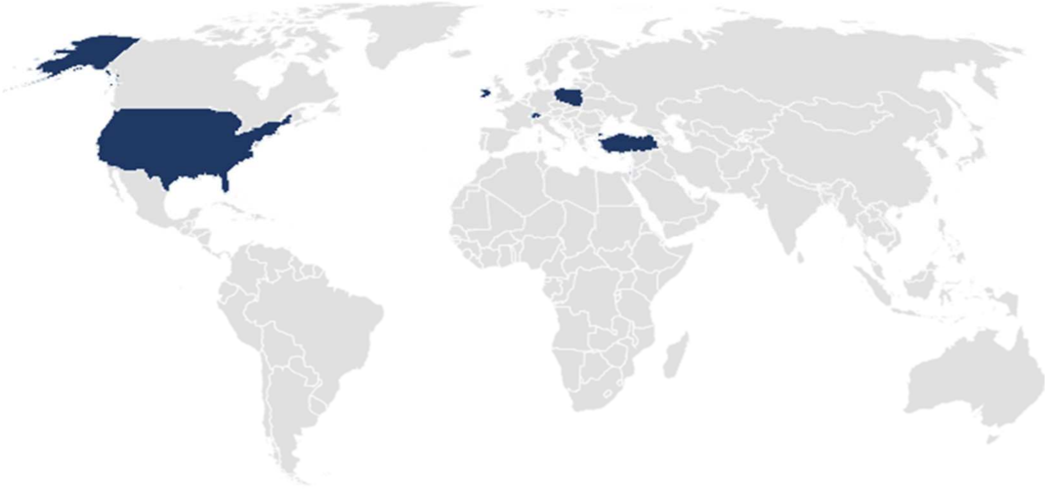


Figure 34 - Map highlighting the countries in which AutoML system considered in the market analysis have been developed

28 tools are owned by companies headquartered in United States, only 7 tools are not in US, it means that the 82% of AutoML tools were developed in US. The other 18% of tools were conceived in Europe. From these data we can say that the best know-how and knowledge about AutoML is developed and deployed in US, this phenomenon is caused by the higher amount of investments in AI made in US compared to other countries.

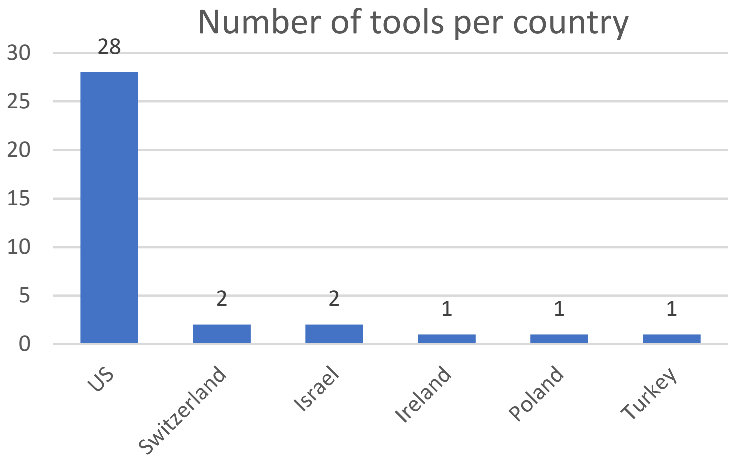


Figure 35 - Number of AutoML tools developed by country

3.5 AutoML Customers Analysis

A very important part of the market analysis is focused on the analysis of the AutoML tools' customer base. For the main tools were searched all the possible confirmed clients that had or are exploiting AutoML services. The aim of this section is to map and to understand in which industries AutoML is more exploited and who are the target customers for each provider. Finally, we will define the market trend of the main industries that are exploiting the AutoML software.

In developing market research, the critical point was researching and validating the customers of AutoML solutions. In order to validate each customer were analyzed all the available use cases shared by AutoML providers on their website. Many use cases found were anonymous and they were not taken into consideration for the analysis.

The research was developed with following steps:

- Choice of the more representative AutoML providers in the market
- Deep research about their customers
- Ensuring that customers found used really AutoML services and no other service of traditional ML through the detailed analysis of the available use cases and report made by the provider or by the customers.
- Collect all the customers in an Excel file with the following features:
- Name of the company
- Industry

AutoML providers for which are available use cases and customers' identity are: Aible, Alteryx, Auger.AI, AWS Sagemaker, Azure ML, Big Squid, Clarifai, Compellon, Dataiku, DataRobot, DMWay, dotData, Featuretools, Google AutoML, H2O Driverless AI, KNIME, Neuton, RapidMiner, SAP Leonardo, SAS.

We chose this list of 20 providers because they share a higher number of information about their customers, instead for other providers there is not public information on their customers. The following table shows all the customers of the previous listed company providing AutoML services:

Provider	Link	Client	Client’s industry
Aible	https://www.aible.com/aible-in-action/	Haas School at UC Berkeley	Education
Aible		Beddr	IT services
Aible		Merrow Sewing Machine	Manufacturing
Aible		Laudio	Software
Alteryx	https://www.alteryx.com/community/customers?industry=All&field_region_target_id=All&title=	Audi	Automotive
Alteryx		Coca-Cola	Consumer goods
Alteryx		Kroger	Retail
Alteryx		Cisco	Technology
Auger.AI	https://auger.ai/	reveal why	Data & Analytics
Auger.AI		YTZ	IT services
Auger.AI		Next	Retail
Auger.AI		SigParser	Software
Auger.AI		frida.ai	Software
AWS SageMaker	https://aws.amazon.com/it/sagemaker/customers/	Korean Air	Airline
AWS SageMaker		Hotels.com	Software
AWS SageMaker		Formosa Plastics	Chemicals
AWS SageMaker		Pioneer	Consumer goods
AWS SageMaker		Dely	Consumer goods
AWS SageMaker		Kinect energy group	Energy industry
AWS SageMaker		Frame.io	Software
AWS SageMaker		Intuit	Financial service
AWS SageMaker		Statefarm	Financial service
AWS SageMaker		Liberty Mutual insurance	Financial service
AWS SageMaker		Coinbase	Financial service
AWS SageMaker		SIGNATE	Graphic company
AWS SageMaker		GE Healthcare	Healthcare
AWS SageMaker		Change Healthcare	Healthcare
AWS SageMaker		Thomson Reuters	Information
AWS SageMaker		SmartNews	Information
AWS SageMaker		ProQuest	Information
AWS SageMaker		Cookpad	IT services
AWS SageMaker		FreakOut	Marketing
AWS SageMaker		terragon group	Marketing
AWS SageMaker		Dow Jones	News and publishing
AWS SageMaker		Celgene	Pharmaceutical
AWS SageMaker		INTERCOM	Software
AWS SageMaker		Tinder	Software
AWS SageMaker		Grammarly	Software
AWS SageMaker		Realtor.com	Software
AWS SageMaker		edmunds.com	Software
AWS SageMaker		Zendesk	Software
AWS SageMaker		Zocdoc	Software
AWS SageMaker		NFL	Sport
AWS SageMaker		F1	Sport
AWS SageMaker		DigitalGlobe	Technology
AWS SageMaker		Siemens	Technology
AWS SageMaker		Regit	Technology
AWS SageMaker		Expedia group	Software
AWS SageMaker		Convoy	Trucking software
Azure	https://azure.microsoft.com/it-it/services/machine-learning/	Schneider Electric	Energy industry
Azure		bp	Energy industry
Azure		TAL	Manufacturing
Azure		Asos	Retail
Azure		walgreens boots alliance	Retail
Azure		Wipro	Technology

Table 13 - Table listing all the AutoML customers, regarding the tools analyzed in the market analysis (continue in next page)

Provider	Link	Client	Client's industry
Big Squid	https://www.bigsquid.com/	Beyond12	Education
Big Squid		USC Suzanne Dworak-Peck	Education
Big Squid		Layton	Energy industry
Big Squid		ENDEAVOR	Financial service
Big Squid		Xyngular	Healthcare
Big Squid		european wax center	Healthcare
Big Squid		UnitedHealth Group	Healthcare
Big Squid		continuum IT mgt platform	IT services
Big Squid		USCCA	Legal protection
Big Squid		Goodwill	Retail
Big Squid		UNTUCKit	Retail
Big Squid		WOMPLY	Software
Big Squid		Skullcandy	Technology
Clarifai	https://www.clarifai.com/customers	Staples	Consumer goods
Clarifai		i-Nside	Healthcare
Clarifai		west elm	Real Estate
Clarifai		Foap	Software
Clarifai		Pixide	Software
Clarifai		Buttercam	Software
Clarifai		Asset Bank	Software
Clarifai		Picturepark	Software
Clarifai		9GAG	Software
Clarifai		Momio	Software
Clarifai		Photobucket	Software
Clarifai		openTable	Software
Clarifai		Vintage cloud	Technology
Compellon	https://www.compellon.com/	The chapman group	Consultancy
Compellon		Andrew Reise	Consultancy
Compellon		Cognizant	IT services
Compellon		1800 contacts	Retail
Compellon		Qlik	Software
Compellon		Clarabridge	Software
Dataiku	https://www.dataiku.com/company/customers/	Tires les schwab	Automotive
Dataiku		evonik power to create	Chemicals
Dataiku		Capgemini	Consultancy
Dataiku		Sephora	Consumer goods
Dataiku		L'Oreal	Consumer goods
Dataiku		Unilever	Consumer goods
Dataiku		Essilor	Consumer goods
Dataiku		FOX networks group	Entertainment
Dataiku		Santander	Financial service
Dataiku		BNP Paribas	Financial service
Dataiku		Premera	Healthcare
Dataiku		OVH.com	IT services
Dataiku		Palo Alto Networks Inc	IT services
Dataiku		dentsu Aegis network	Marketing
Dataiku		Nuxeo	Software
Dataiku		Sendinblue	Software
Dataiku		Callidus Cloud	Software
Dataiku		Ubisoft	Software
Dataiku		KUKA	Technology
Dataiku		COMCAST	Telco
DataRobot	https://www.datarobot.com/success/customers/	virgin Australia	Airline
DataRobot		united airlines	Airline
DataRobot		O-BASF	Chemicals
DataRobot		Deloitte	Consultancy

Table 13 - Table listing all the AutoML customers, regarding the tools analyzed in the market analysis (continue in next page)

Provider	Link	Client	Client’s industry
DataRobot	https://www.datarobot.com/success/customers/	Accenture	Consultancy
DataRobot		Michigan university	Education
DataRobot		Crest Financial	Financial service
DataRobot		ledingtree	Financial service
DataRobot		usbank	Financial service
DataRobot		Symphony Post Acute	Healthcare
DataRobot		Steward	Healthcare
DataRobot		Humana	Healthcare
DataRobot		New york life	Financial service
DataRobot		Bluecross BlueShield	Financial service
DataRobot		Aegon	Financial service
DataRobot		Torque Data	Marketing
DataRobot		One Marketing	Marketing
DataRobot		kroger	Retail
DataRobot		Carrefour	Retail
DataRobot		tableau	Software
DataRobot		Traveloka	Software
DataRobot		Philadelphia 76ers	Sport
DataRobot		panasonic	Technology
DataRobot		Lenovo	Technology
DMway	featuredcustomers.com/vendor/dmway/testimonials	Forbes	Information
DMway		Data science central	Data & Analytics
DMway		Ono academic college	Education
DMway		Ituran	Logistics
DMway		Gartner	Research, consulting
DMway		Red Herring	Technology
dotData	https://dotdata.com/white-papers/	Japan Airlines	Airline
dotData		SMBC Group	Financial service
dotData		MS&AD Insurance group	Financial service
Featuretools	https://www.featuretools.com/	Accenture	Consultancy
Featuretools		MIT	Education
Featuretools		BBVA	Financial service
Featuretools		DARPA	Technology
Google AutoML	https://cloud.google.com/automl/?hl=it	Chevron	Energy industry
Google AutoML		Disney	Entertainment
Google AutoML		Imagia	Healthcare
Google AutoML		Meredith Digital	Marketing
Google AutoML		URBN	Retail
Google AutoML		California Design Den	Retail
H2O Driverless AI	https://www.h2o.ai/customer-stories/	Ducit.ai	Financial service
H2O Driverless AI		Deserve	Financial service
H2O Driverless AI		Vision Banco	Financial service
H2O Driverless AI		Dun & Bradstreet	Financial service
H2O Driverless AI		Equifax	Financial service
H2O Driverless AI		Paypal	Financial service
H2O Driverless AI		Capital One	Financial service
H2O Driverless AI		ADP	Financial service
H2O Driverless AI		pwc	Financial service
H2O Driverless AI		Wells Fargo	Financial service
H2O Driverless AI		Underwrite.ai	Financial service
H2O Driverless AI		ArmadaHealth	Healthcare
H2O Driverless AI		Reproductive science center	Healthcare
H2O Driverless AI		Change Healthcare	Healthcare
H2O Driverless AI		HCA	Healthcare
H2O Driverless AI		Kaiser permanente	Healthcare
H2O Driverless AI		AEGON blue square re	Financial service

Table 13 - Table listing all the AutoML customers, regarding the tools analyzed in the market analysis (continue in next page)

Provider	Link	Client	Client’s industry
H2O Driverless AI	https://www.h2o.ai/customer-stories/	Zurich insurance	Financial service
H2O Driverless AI		Progressive insurance	Financial service
H2O Driverless AI		ING	Financial service
H2O Driverless AI		Macnica Networks	Manufacturing
H2O Driverless AI		Hortifrut	Manufacturing
H2O Driverless AI		Stanley Black & Decker	Manufacturing
H2O Driverless AI		Intel	Manufacturing
H2O Driverless AI		Marketshare	Marketing
H2O Driverless AI		Beeswax	Marketing
H2O Driverless AI		Nielsen catalina solution	Marketing
H2O Driverless AI		G5	Marketing
H2O Driverless AI		Macy's	Retail
H2O Driverless AI		Booking.com	Software
H2O Driverless AI		Travelport	Software
H2O Driverless AI		Comcast	Telco
H2O Driverless AI		Tech Mahindra	Telco
KNIME	https://enlyft.com/tech/products/knime	Procter & Gamble	Consumer goods
KNIME		Harnham	Data & Analytics
KNIME		Horizon Media	Entertainment
KNIME		United community Bancorp	Financial service
KNIME		Morgan Stanley	Financial service
KNIME		Ironwood Pharmaceutical	Healthcare
KNIME		Palo Alto Networks Inc	IT services
KNIME		National Fire Protection	Safety
KNIME		ConvertCart	Software
KNIME		Tyler technologies	Software
Neuton	http://bellintegrator.com/Neuton	zepter international	Consumer goods
Neuton		shell	Energy industry
Neuton		CityBank	Financial service
Neuton		societe generale	Financial service
Neuton		Deutsche Bank	Financial service
Neuton		juniper networks	Technology
Neuton		cisco	Technology
Neuton		CenturyLink	Telco
Neuton		ericsson	Telco
RapidMiner	https://rapidminer.com/case-studies/	LIAT	Airline
RapidMiner		Lufthansa	Airline
RapidMiner		BMW	Automotive
RapidMiner		TfL	Chemicals
RapidMiner		GE	Conglomerate
RapidMiner		Miele	Consumer goods
RapidMiner		Mobilkom	Entertainment
RapidMiner		Paypal	Financial service
RapidMiner		Daimler	Manufacturing
RapidMiner		cisco	Technology
RapidMiner		Samsung	Technology
RapidMiner		Body Biolytics	Technology
RapidMiner		Intel	Technology
RapidMiner		SustainHub	Value chain
SAP Leonardo	https://www.sap.com/products/analytics/predictive-analytics.html	Lloyd's Register	Energy industry
SAP Leonardo		Groupe Mutuel	Financial service
SAP Leonardo		Office of State Revenue in	Government
SAP Leonardo		DuluxGroup	Manufacturing
SAS	https://www.sas.com/it_it/software/visual-data-mining-machine-learning.html	Seacoast Bank	Financial service
SAS		UMC Utrecht	Healthcare
SAS		Amsterdam UMC	Healthcare

Table 13 - Table listing all the AutoML customers, regarding the tools analyzed in the market analysis

As we can see from the table for each customer was featured with the belonging industry. This research was made with the purpose of mapping the customer base of each tool.

Analyzing AutoML clients and classifying them we obtain a meaningful graph that explains what are the main industries that are leveraging on AutoML services.

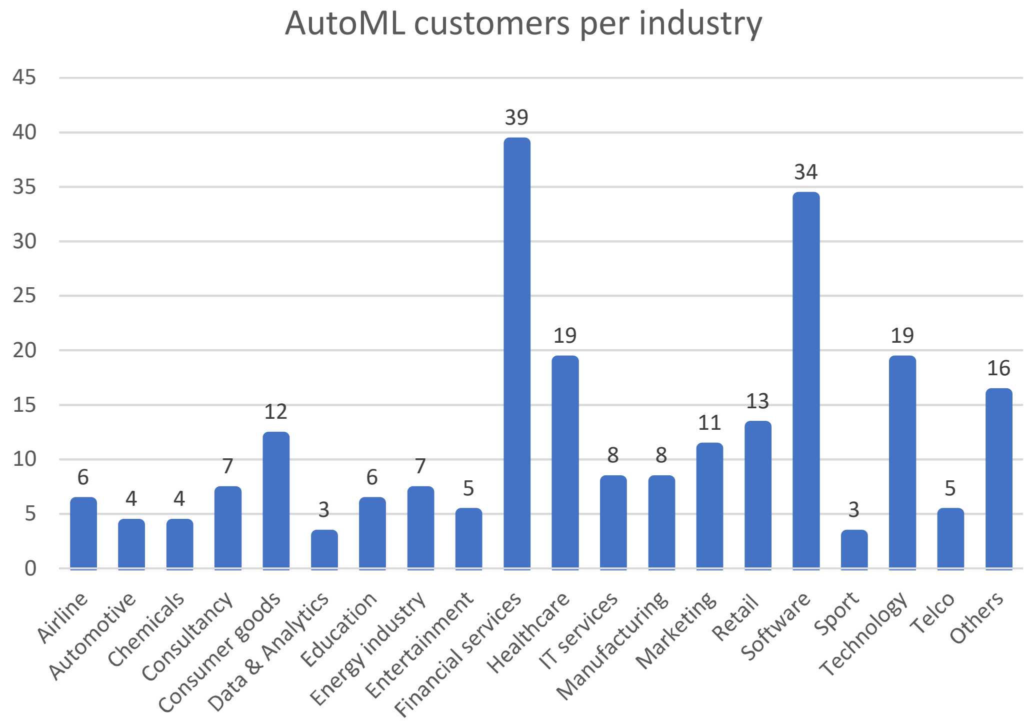


Figure 36 – Count of AutoML customers per industry

As we can see there are two main trends: in Financial Services (17.03%) and in Software industries (14.84%).

Thank the graph is easy to notice that the implementation of AutoML solutions is still restricted to some specific businesses because those industries are leveraging on data more than others and they need technologies and competencies to improve their performances. Always from the histogram, we can notice that even other industries are leveraging quite a lot on AutoML systems, for example, Technology and Healthcare industries have a consistent number of use cases in which AutoML is exploited.

Now that the trend of industries using AutoML solutions is defined, the following part of the market research will focus on the type of company that needs AutoML, the objective is to understand if only big companies leverage on AutoML or even small\medium companies use it. To do this was collected for each customer its size, considering the number of employees actually working for them. Data about the size were found both on Linkedin, Crounchbase, and Wikipedia. In doing this analysis we consider the following company classification:

- Small companies: when the number of employees < 50
- Medium companies: when the number of employees is in the range 50 - 250
- Big companies: when the number of employees > 250

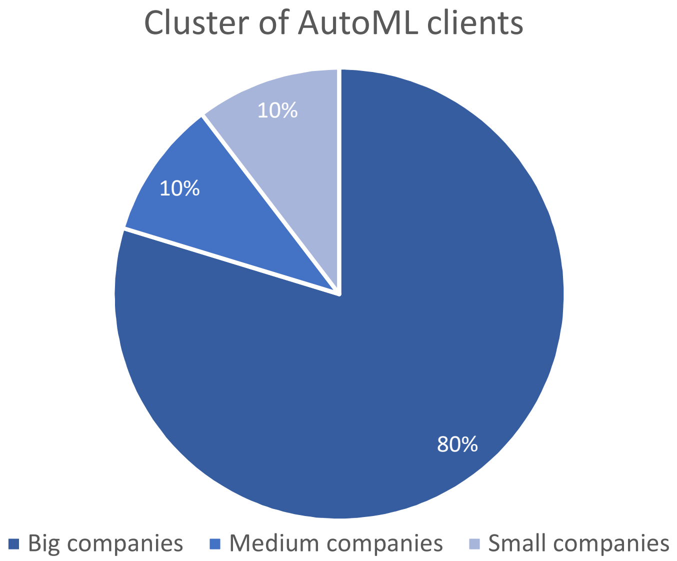


Figure 37 - Pie chart to highlight the classes of AutoML customers considering their size

From the pie chart is evident that AutoML systems have a strong presence in big companies, indeed 80% of customers are big companies. The remaining 20% is perfectly split, 10% of small companies and 10% of medium companies. The reasons of this segmentation are in the following points:

- Big companies leverage on AutoML to increase the productivity of ML process, to save time and money, to save human resources and to create a data-driven culture inside the company in each business unit, enabling everyone taking decision and dealing with data. Since big companies must manage every day Big Data these tools are helpful to process and analyze the data with less effort and automatic. Furthermore, big companies often tend to invest a lot in new technologies because they have the budget to invest and then also to create a competitive advantage over their competitors.
- Small/medium companies leverage on AutoML systems to fill the need for analytics competencies, with the consequences of increasing the efficiency of internal and external processes. AutoML is used also to enter new markets having a good data-driven approach that enables the company to take important decisions driven by data and not yet by assumptions.

For both there are two common benefits, one is the possibility to save time and money in developing a ML project, and second to fill the gap of ML experts available today.

Now we have a picture of the main industries using AutoML and the distribution of companies using AutoML based on their size. The next and last step is to build a picture showing for the main AutoML tools what are their customer base characteristics. To visualize the actual situation were calculated the average size of the customer base of each AutoML tool and then created the following histogram that compare the different target of each system:

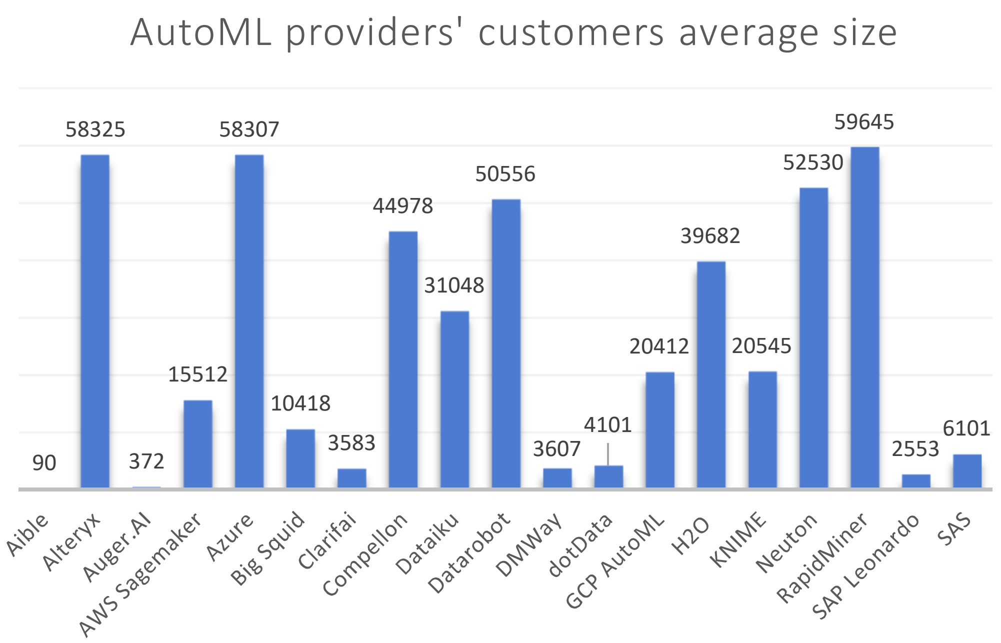


Figure 38 - Bar chart highlighting the average size of the customers per AutoML tool

The graph highlights that each tool has its own specific target of customers.

Considering the division made to classify the company size, what emerged is that the general targeting is toward big companies, with the exception of Aible that shows a target customer base classified as medium size and also Auger.AI has a lower targeting compared with other providers.

As we can see there are 8 providers that have an average customer base size higher than 30,000 employees, they serve very large companies.

We can conclude from the whole analysis about customers that today AutoML is exploited mainly in big companies and that there are two main industries in which it is used: financial services and software.

3.6 Deep-dive on AutoML tools used in business case

This section aims to make a deep analysis and description of every AutoML tool used in the business case in Chapter 4 to assess their characteristics. The methodology followed to analyze the tools was defined as a benchmark framework that could be applied to every AutoML tool to assess its characteristics, the idea of the benchmark born from the academic research DataBench, 2018:

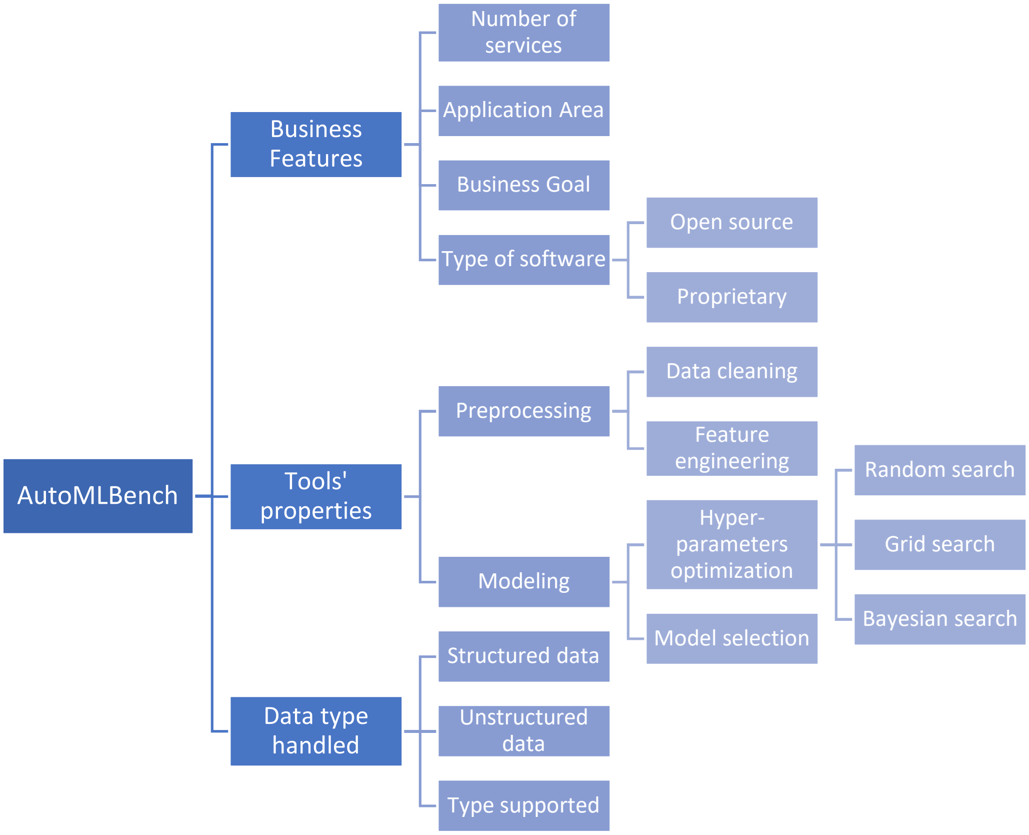


Figure 39 - Framework developed to analyze in depth an AutoML tool

Every tool was analyzed under 3 main characteristics: Business Features, Tool's Properties and Data Type Handled. The framework has the aim to help in assessing the characteristic of each AutoML software and in comparing the different solutions with the same framework. It is a 'high level' comparison tool, in fact the usage of this benchmark is useful to build an idea of what kind of services one system may perform. For each tool tested in Chapter 4 there is a dedicated table built from Figure 36. All the information to build the following table are extracted from the official documentation of each AutoML systems.

Google Cloud AutoML		
Business Features	Software info	Launch date January 2018, still working in Beta. It is classified as a Generalized Proprietary tool.
	Services	Google provides 5 different services: Google AutoML Tables, Google AutoML Video Intelligence, Google AutoML Vision, Google AutoML Natural Language, Google AutoML Translation.
	Application Area	Customer service and support, engineering, maintenance and logistics, marketing, finance, sales, IT and data operations.
	Use cases	Price optimization, inventory and service parts optimization, product & service recommendation system, fraud prevention and detection, customer profiling, targeting and offers optimization.
Technical Properties	Preprocessing	AutoML Tables helps you create clean, effective training data by providing information about missing data, correlation, cardinality, and distribution for each of your features. AutoML Tables automatically performs common feature engineering tasks for you, including: normalize and bucketize numeric features, create one-hot encoding and embeddings for categorical features, perform basic processing for text features, extract date- and time-related features from Timestamp columns.
	Modeling	When you kick off training for your model, AutoML Tables takes your dataset and starts training for multiple model architectures at the same time. As new model architectures come out of the research community, we will add those as well.
Data Type Handled	Structured	Tabular data, time series.
	Unstructured	Images, Videos, Audios, Texts.
Algorithms	Linear, feedforward deep neural network, Gradient Boosted Decision Tree, AdaNet, Ensembles of various model architectures.	
Further details	https://cloud.google.com/automl/docs/	

Azure ML Studio Microsoft		
Business Feature	Software info	It is a generalized proprietary tool.
	Services	Azure ML Studio (classification, regression, time series).
	Application Area	Customer service and support, engineering, maintenance and logistics, marketing, finance, sales, IT and data operations.
	Use cases	Fraud detection, CPU performance prediction, demand forecasting, marketing prediction, material durability prediction, sales forecasting.
Technical Properties	Preprocessing	In every automated ML experiment, your data is automatically scaled or normalized to help algorithms perform well. During model training, one of the following scaling or normalization techniques will be applied to each model: StandardScaleWrapper, MinMaxScalar, MaxAbsScaler, RobustScalar, PCA, TruncatedSVDWrapper, SparseNormalizer.
	Modeling	Automated ML supports ensemble models, which are enabled by default. Ensemble learning improves ML results and predictive performance by combining multiple models as opposed to using single models. The ensemble iterations appear as the final iterations of your run. Automated Machine Learning uses both voting and stacking ensemble methods for combining models.
Data Type Handled	Structured	Tabular data, time series.
	Unstructured	Texts, images, video.
Algorithms	• Two-class classification: logistic regression, decision forest, decision jungle, boosted decision tree, neural network, averaged perceptron, support vector machine, locally deep support vector machine, Bayes' point machine. • Multi-class classification: logistic regression, decision forest, decision jungle, one-v-all. • Regression: linear, Bayesian linear, decision forest, boosted decision tree, fast forest quantile, neural network, Poisson, ordinal. • Anomaly detection: support vector machine, PCA-based, K-means.	
Further details	https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-automated-ml	

Dataiku		
Business Feature	Company info	It was founded on February 14, 2013 in Paris (FR). The Data Science Studio (DSS) was released in 2014, Dataiku is classified as a Generalized Proprietary tool. It is headquartered in New York City (US) and has 200 employees.
	Number of services	Data Science Studio (DSS)
	Application Area	Marketing analytics, logistics analytics, R&D analytics, business intelligence, data labs, sales analytics, human resource analytics.
	Use cases	Churn analytics, fraud detection, graph analytics, data management, demand forecast, spatial analytics, lifetime value optimization, predictive maintenance, analytical CRM
Technical Properties	Preprocessing	Automatic features engineering, generation and selection to use any kind of data in your models.
	Modeling	Optimize your model hyperparameters using various cross validation strategies. Compare dozens of algorithms from Dataiku interface, both for supervised and unsupervised tasks. Get instant visual insights from your model (variables importance, features interactions or parameters), and assess model's performance through detailed metrics.
Data Type Handled	Structured	Tabular, time series
	Unstructured	Images, videos and texts
Algorithms	• Python-based: Ordinary Least Squares, Ridge Regression, Lasso Regression, Logistic regression, Random Forests, Gradient Boosted Trees, XGBoost, Decision Tree, Support Vector Machine, Stochastic Gradient Descent, K Nearest Neighbors, Extra Random Trees, Artificial Neural Network, Lasso Path, Custom Models offering scikit-learn compatible API's (ex: LightGBM). • Spark MLLib-based: Logistic Regression, Linear Regression, Decision Trees, Random Forest, Gradient Boosted Trees, Naive Bayes, Custom models. • H2O-based: Deep Learning, GBM, GLM, Random Forest, Naive Bayes.	
Further details	https://doc.dataiku.com/dss/latest/	

AWS Sagemaker		
Business Feature	Company info	It was released on November 29, 2017. It is a narrow proprietary tool used only for automating the hyper-parameters setting.
	Number of services	Hyperparameters tuning
	Application Area	Marketing analytics, logistics analytics, R&D analytics, business intelligence, data labs, sales analytics.
	Use cases	Recommendation, forecasting, image and video analysis, advanced text analytics, document analysis, voice, conversational agents, translation, transcription,
Technical Properties	Preprocessing	Null
	Modeling	AWS Sagemaker only automate the hyper-parameters tuning, it tunes the hyper-parameters of the model selected manually. For each hyper-parameters is possible to set the range of values it can assume.
Data Type Handled	Structured	Tabular, time series
	Unstructured	Image, video and text
Algorithms	• Random Search In a random search, hyperparameter tuning chooses a random combination of values from within the ranges that you specify for hyperparameters for each training job it launches. • Bayesian Search Bayesian search treats hyperparameter tuning like a [regression] problem. Given a set of input features (the hyperparameters), hyperparameter tuning optimizes a model for the metric that you choose. To solve a regression problem, hyperparameter tuning makes guesses about which hyperparameter combinations are likely to get the best results and runs training jobs to test these values. After testing the first set of hyperparameter values, hyperparameter tuning uses regression to choose the next set of hyperparameter values to test.	
Further details	https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html	

H2O Driverless AI		
Business Feature	Company info	It was launched on 24 September 2017. It is classified as a generalized open source tool.
	Number of services	H2O Driverless AI
	Application Area	Financial services, insurance, healthcare, marketing, telecom, manufacturing, retail
	Use cases	Advance analytics, fraud detection, claims management, digital advertising.
Technical Properties	Preprocessing	Data that is imported into Driverless AI can include missing values. Feature engineering is fully aware of missing values, and missing values are treated as information - either as a special categorical level or as a special number. So, for target encoding, for example, rows with a certain missing feature will belong to the same group. Driverless AI will automatically do variable standardization for certain algorithms. For example, with Linear Models and Neural Networks, the data is automatically standardized. For decision tree algorithms, however, we do not perform standardization since these algorithms do not benefit from standardization. Also, features are engineered with a proprietary stack of Kaggle-winning statistical approaches including some of the most sophisticated target encoding and likelihood estimates based on groupings, aggregations and joins, but we also employ linear models, neural nets, clustering and dimensionality reduction models and many traditional approaches such as one-hot encoding etc
	Modeling	No specified method
Data	Structured	Tabular, time series.
	Unstructured	It does not support image/video/audio
	Type supported	arff, bin, bz2, csv, dat, feather, gz, jay, nff, parquet, pkl, tgz, tsv, txt, xls, xlsx, xz, zip
Algorithms	XGBoost, LightGBM, GLM, Tensorflow, RuleFit, FTRL.	
Further details	http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html	

4 BUSINESS CASE: A COMPARISON BETWEEN AUTOML AND ML MODELS

In this chapter, the focus is on the study and the analysis of a real case developed in Bip. The case is very useful to make sense at all the previous chapters because it touches all the topics treated: AutoML tools, impact on the organization and performance between the traditional ML and AutoML. The business case has the objective to compare the performances obtained by the traditional ML and the performances obtained by AutoML systems, the problem treated is about a prevision model developed by Bip for its client in the previous years. All the experiments on different AutoML tools were executed in Bip.xTech on datasets used to create the real ML prevision model for the client.

The purpose of writing this business case is to test and to evaluate the AutoML tools' performances when dealing with a forecasting problem and to highlight the benefits and constraints these tools have in real implementations. For these reasons were tested five different AutoML systems on the same datasets to understand the potentialities of each one and at the same time to compare the performances between them in reaching the same goal.

This chapter will be structured in the following way:

- Introduction to the business of the company for whom Bip developed the ML projects.
- Regulatory problem: the reason why this company needed a very accurate forecasting model.
- Description of the first solution developed by Bip (2017), the Artificial Neural Network (ANN).
- Description of the second solution developed by Bip (2018), it is an Ensemble of 6 sub-models where every model gives a contribution to the final prediction, we will refer to it with the alias (Ensemble)
- AutoML experiments to solve the same problem using 5 different tools.

To have a big picture about the evolution of our business case the following image explains the path followed to achieve the actual prediction model (Ensemble) developed in Bip for a gas infrastructure company:

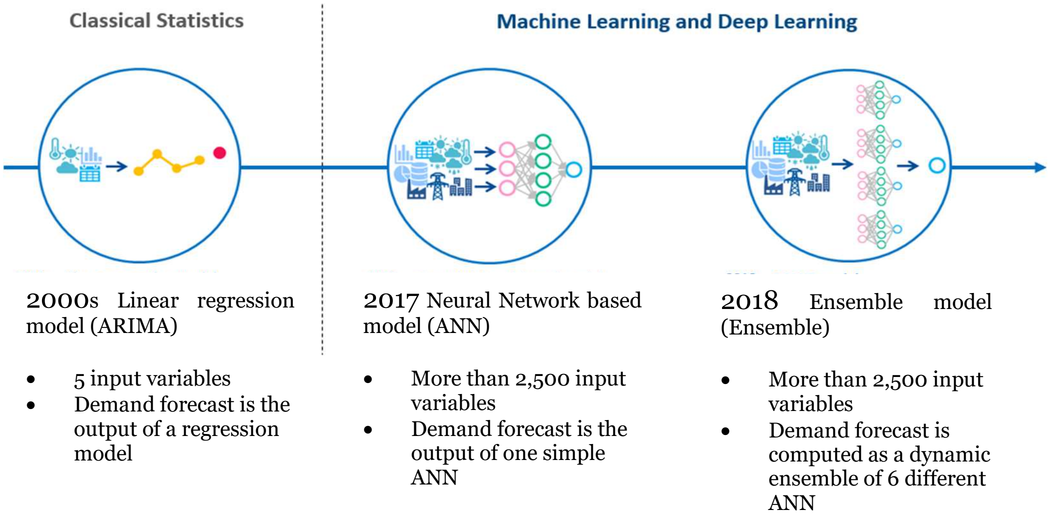

Classical Statistics
Machine Learning and Deep Learning
2000s Linear regression model (ARIMA)
5 input variables
Demand forecast is the output of a regression model
2017 Neural Network based model (ANN)
More than 2,500 input variables
Demand forecast is the output of one simple ANN
2018 Ensemble model (Ensemble)
More than 2,500 input variables
Demand forecast is computed as a dynamic ensemble of 6 different ANN

Figure 40 - Evolution of ML models to predict the gas day-ahead

The gas company until 2017 applied a simple statistical model to make predictions about gas consumption. Bip improved the performance of predictions with an artificial neural network (ANN) model and then with the best performing model named Ensemble, an ensemble of 6 sub-models.

Before talking about ML and AutoML models is necessary to define the client's business context and its problems. As we will see in detail in the next section the problem Bip was asked to solve was to create a more precise predictive model because the model in production before 2017 did not allow the client to gain incentives from a regulatory law.

Then we will introduce firstly the ANN model highlighting the differences with the ARIMA model and then we will introduce the Ensemble model highlighting the differences with the ANN model.

4.1 Business context

The business case used to assess the differences between ML and AutoML is related to a prediction problem for the daily distribution of gas in Italy. The name of the company must be censured for privacy reasons.

The client is a big company that works in the industry of Oil & Gas and it is present in Europe as one of the leading companies. Its role consists in receiving gas from producers or shippers, transporting it via pipeline, and delivering it to second level gas distribution companies or directly to industries and power plants.

The final goal of this company is the gas dispatching, it is a Transmission System Operator (TSO). The phase of dispatching is composed by four main processes and the company must manage them very carefully:

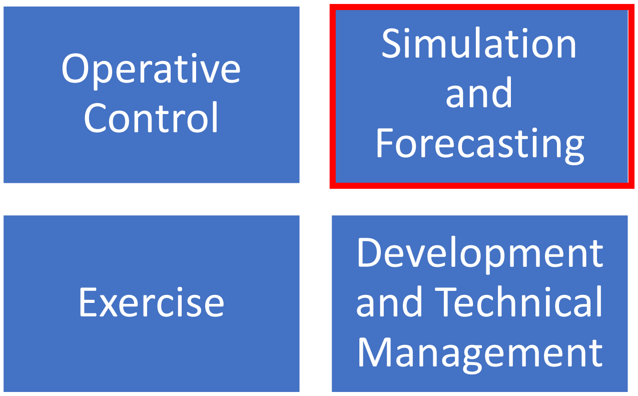

Operative Control
Simulation and Forecasting
Exercise
Development and Technical Management

Figure 41 - Four main processes of gas dispatching

About what concerns the business case, we will focus on the process of Simulation and Forecasting because to solve the regulatory problem, better introduced in the next section, we need to improve the process of forecasting.

This is what concern the core business of this company, now with the following section will be explained why this company needed a very accurate forecasting model.

4.2 Regulatory problem

The problem was born when was introduced a new regulation law for the gas dispatching. The regulatory process of the Gas Value Chain was born in 1998 in Europe. The objective of the regulation is to guarantee access to value chain services and the conditions parity for each user. The European regulation impacts and defines the Italian regulation under which the company operates.

The main activities regulated are:

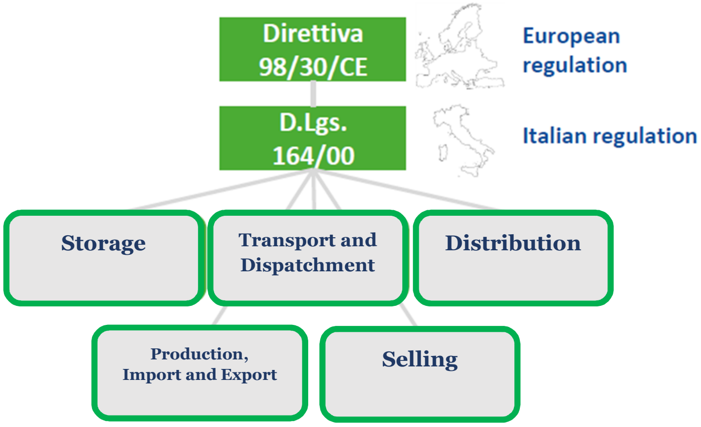


Figure 42 - Legislation structure

The system of balancing is regulated by the European normative EU 312/2014, that has the following objectives:

- To create a market completely operative and interconnected.
- To ensure the provision at lower prices.
- To increase the competitiveness and the market liquidity.
- To increase the gas provision flexibility.

The directive UE 312/2014 is applied in Italy by resolution 312/2016/R/GAS, the resolution has 2 main principles:

Resolution	Users	Evaluation frequency
Commercial balancing	The responsibility of the network users lies with to guarantee equality the input and the withdrawal required for correct accounting and allocation of the transported gas.	Daily
Physical balancing	The transporter must check the flow parameters in order to guarantee every time the safe and correct movement of the gas from the entry points to the withdrawal points. The regulation states that "users are responsible for balancing their balance sheet portfolios so as to minimize the need for transport system operators to take balancing actions.	Real time

Table 14 - Principles of the resolution

What is important for the company is to analyze the imminent future to optimize the gas distribution. There are three main type of previsions:

1) **Expected balance of the system**

It is the estimate based on the nominations of users to promote actions on the market of rebalancing the system.

2) **Redelivered forecast**

It represents the forecast of the gas taken by users at redelivery points turn the day G1 for day G through a ML model.

3) **Line-Pack forecast**

It represents the forecast of gas variation in the transport network.

To follow coherent balancing actions with the economic and efficient functioning of the transport network, the authority has instituted three performance indicators of transporter (validated from 27/11/2018):

Performance indicators	Description
P1: re-delivery forecast	It measures the daily forecast error of gas withdrawn, in the precedent day to that of flow. The P1 incentives the company to elaborate forecasts of re-delivery very accurate.
P2: stock price buying/selling	P2 measures the relationship between the difference of stock prices of balancing and the weighted average price of market in every gas day. P2 incentives the company to respect the neutrality of transporter.
P3: residual balancing	P3 measures the usage for the balancing of the resources network in the availability of RdB. P3 incentives the company to limit the usage of his own resources and it fosters the intervention on the market with the provision of products.

Table 15 - Table listing the three performances instituted by the authority

The Italian regulatory authority for energy, networks and the environment (ARERA) introduced an incentive to encourage accurate day-ahead delivered gas forecasts. The explanation of how this incentive works is shown in the following graph:

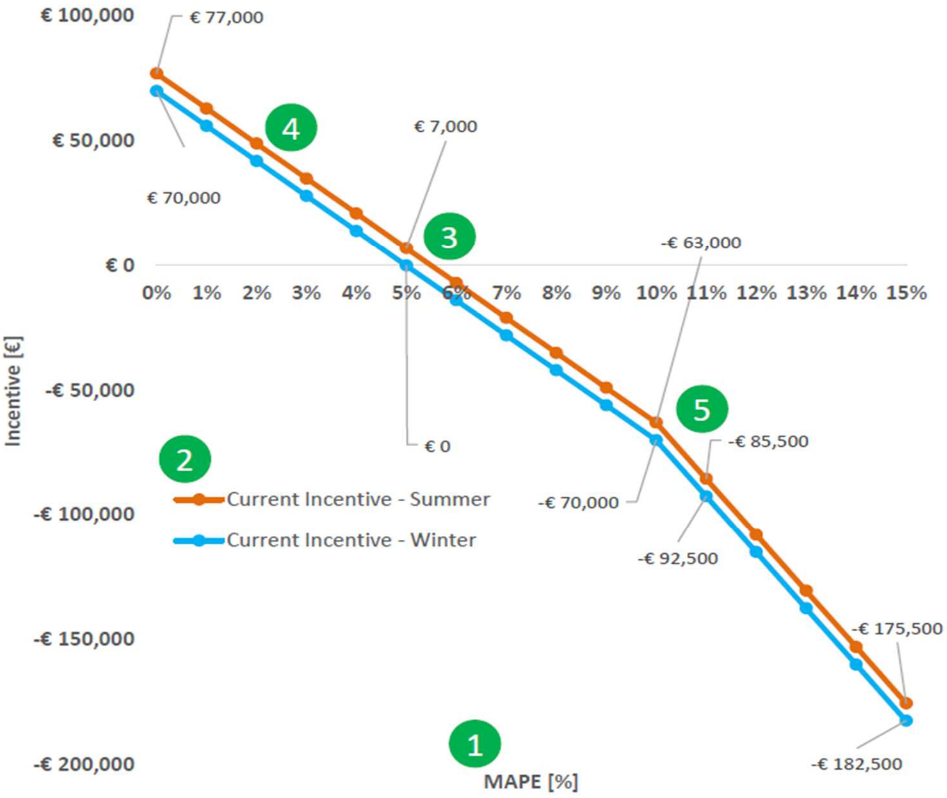


Figure 43 - Incentives structure, relation between MAPE and Incentive

The graph shows the amount of incentive that gas infrastructure company could get or pay according to the relative MAPE for the gas day-ahead forecasting. There are 5 points important to understand the incentive:

1. Performances are measured in terms of Mean Absolute Percentage Error (MAPE), already introduced in chapter 1.4.6, which can be calculated as:

$$\text{MAPE} = \frac{100\%}{n}\sum_{1}^{n}|\frac{Predicted - Actual}{\text{Actual}}|$$

2. Distinction between summer and winter, according the different gas consumption profiles. The goal is to decrease the penalization during the summer period (in which the delivered volumes are low).
   - The summer period refers from April to September.
   - The winter period refers from October to March.
3. Two different profiles correspond to two different null incentive points:
   - Null incentive corresponds to a 5% error in winter.
   - Null incentive corresponds to a 5.5% error in summer.

4. Linear incentive for errors below 10%, where 1% variation of MAPE corresponds to a monetary variation of 14,000 €.
5. Increment of the penalization for errors above 10%, where a 1% variation of MAPE corresponds to a monetary variation of 22,500 €. The goal is to increase the accuracy of predictions, prioritizing the reduction of the volatility.

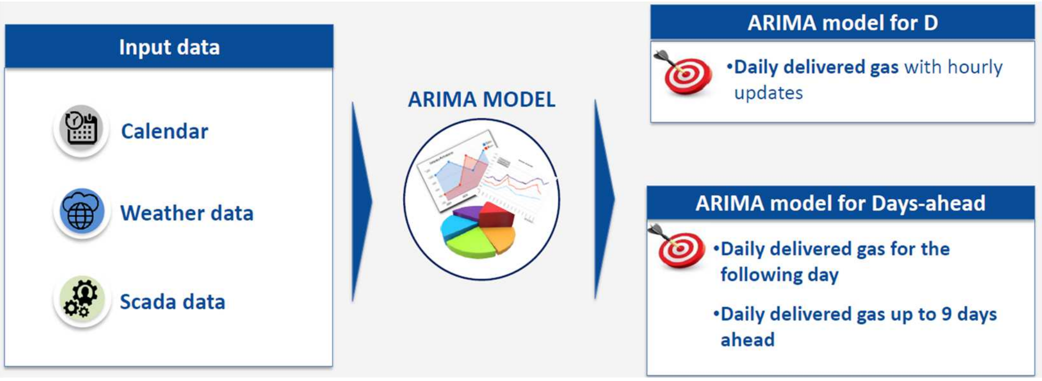
Big government subsidies are given if the prediction is accurate enough to favor the competition on the free market of energy: if the prediction has an error below 5% (during winter, in summer the null incentive is 5.5%) then the company receives subsidies, otherwise the company must pay a fine.

This new regulation created at the same time both an opportunity and a real dangerous situation for the gas company. To exploit the regulation as an opportunity it requested to Bip to develop an efficient and accurate prediction model for the gas day-ahead.

4.3 Situation before Bip

Before Bip, the gas company applied an old methodology to predict the gas-day consumption. Its method was the ARIMA, a statistics method that considered few variables. ARIMA is considered a classic statistics method, it was implemented since 1980s until 2017 when Bip developed its first ML model to predict in a better way the gas consumption.

ARIMA that stand for AutoRegressive Integrated Moving Average is a linear autoregressive model, fitted to delivered gas time series data. It got in input just three kind of data to create forecasts: Calendar, Weather Data and Scada Data. Then the model releases a forecast for D and forecasts for until 9 days ahead.



Input data

Calendar

Weather data

Scada data

ARIMA MODEL

ARIMA model for D

•**Daily delivered gas** with hourly updates

ARIMA model for Days-ahead

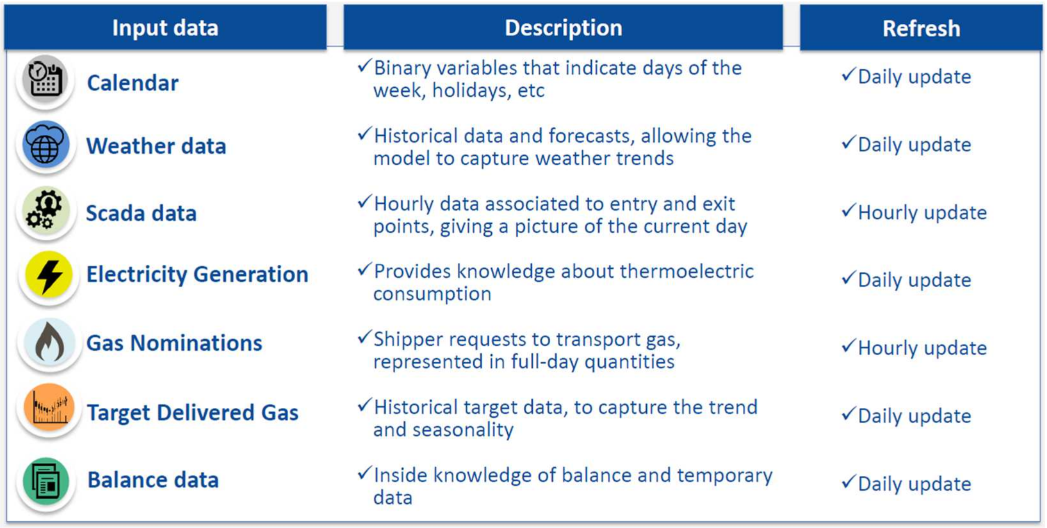
•**Daily delivered gas for the following day**

•**Daily delivered gas up to 9 days ahead**

Figure 44 - ARIMA model input and output

The performance of this initial solution will be analyzed in the next section to have a comparison with the results of the first solution developed in Bip.

To understand the type of data in input to the different models was built Table 14. It resumes all the type of data used to create the different models that will be introduce in the next sections.



Input data	Description	Refresh
Calendar	✓Binary variables that indicate days of the week, holidays, etc	✓Daily update
Weather data	✓Historical data and forecasts, allowing the model to capture weather trends	✓Daily update
Scada data	✓Hourly data associated to entry and exit points, giving a picture of the current day	✓Hourly update
Electricity Generation	✓Provides knowledge about thermoelectric consumption	✓Daily update
Gas Nominations	✓Shipper requests to transport gas, represented in full-day quantities	✓Hourly update
Target Delivered Gas	✓Historical target data, to capture the trend and seasonality	✓Daily update
Balance data	✓Inside knowledge of balance and temporary data	✓Daily update

Table 16 - Input data

4.4 Bip first solution: ANN

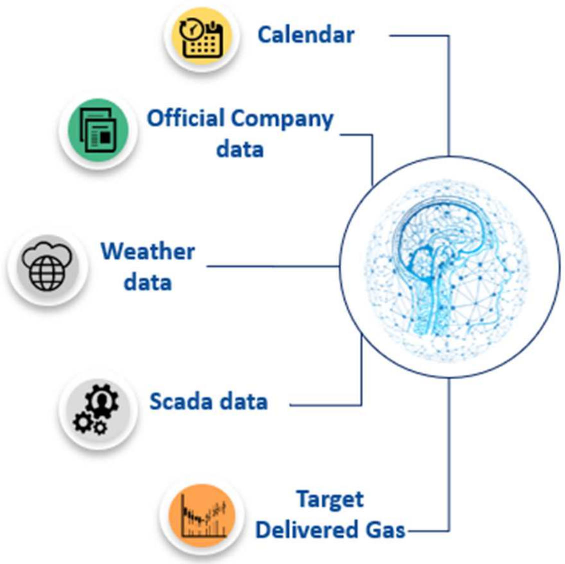


Figure 45 - Data input in ANN

In 2017 Bip was requested to improve the accuracy of predictions for delivered gas in the day-ahead.

Bip developed a model able to provides hourly predictions of the volume of gas delivered, with a yearly average percentage error less than 4%. It is based on a model that consider real time data about weather and gas usage as well as weather forecasts and other official gas company data. The implementation has a rolling operating mode, since the models are continuously retrained with the most recent information. This ML model got in input more data than the previous ARIMA model: Calendar data, Official Company data, Weather data, Scada data and Target Delivered Gas.

The power of ML model is the ability to deal at the same time with many different data sources and exploit the potential of big data, extracting value from them. The input data are all possible variables used to explain the target variable, collected from multiple data sources. Then the algorithm is capable of characterizing the patterns between the input data and the target, providing as output a synthetic business-oriented information. The outputs of the model were:

- Demand forecast for D.
- Daily delivered gas with hourly updates.
- Demand forecast for Day-ahead.
- Daily delivered gas for the following day, with detail per client and hourly updates.
- Trend demand forecast.
- Daily delivered gas up to 4 days ahead, with hourly updates.

Main characteristics of the ANN model are:

Artificial Neural Network
Automated hourly forecasts
Project duration: 24 weeks
2.5 FTE (Full Time Employee)
Prediction error below <4% (MAPE)
+20% improvement of the system w.r.t. baseline
+5M€ of incentives gained in 1 year
Year 2017

Table 17 - ANN main characteristics

To have a clear picture of the improvements brought by ANN the next graphs show the simulation performances (Oct 2016 – Sep 2017) and the production performances (Oct 2017 – Sep 2018) of ANN in comparison with ARIMA.

Simulation performances ARIMA vs ANN (Oct 2016 – Sep 2017)

Simulation were carried out to validate the day-ahead model and compare its performances to the previous model that was running live.

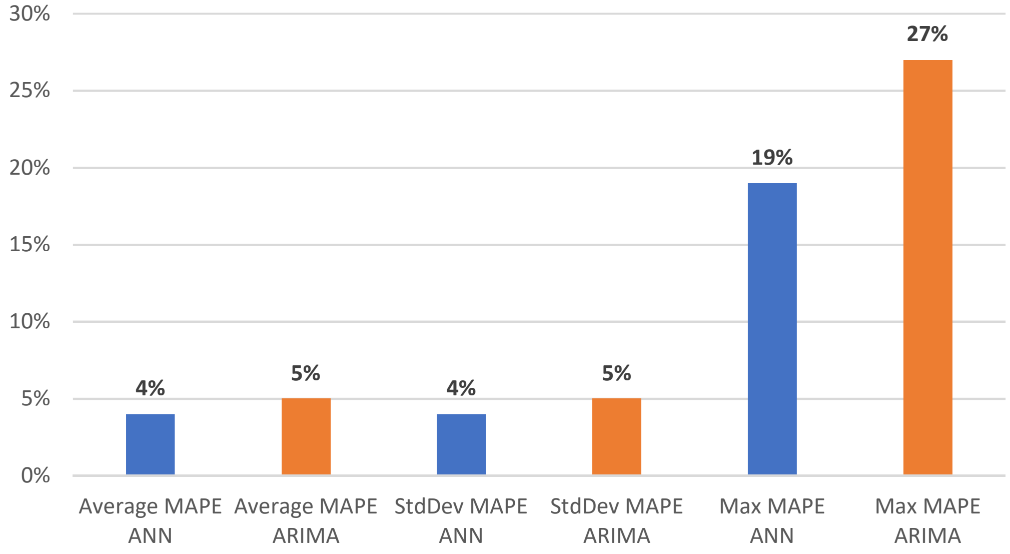


Figure 46 - Comparison between ANN and ARIMA, in simulation

The results of the simulation say that ANN reduce the average error of 1% and also the variability is reduced a lot, since that the maximum MAPE is 19%. Even the standard deviation of the MAPE was better of 1 % and the maximum MAPE obtained during the time period was 8% less than the maximum MAPE obtained by the ANN developed the year before.

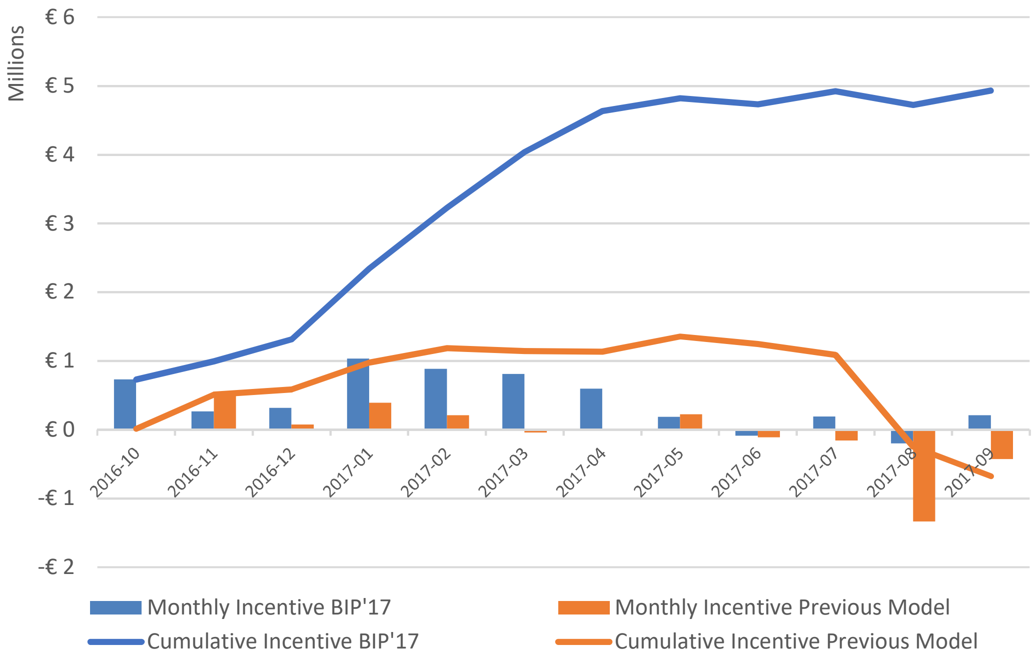


Figure 47 - Incentives comparison between ANN and ARIMA in simulation

Having obtained these optimal results in the simulation, on Oct 1 ANN was released and put in production.

The chart main result is the reduction of the MAPE with the consequence of a gain from incentives equal to ≈ 5M €. Instead with the previous model the simulation resulted with a loss of ≈ 0.5M €. A very good result is that the maximum error is decreased by 8%, this mean that the model works better than the previous and avoid critical situation as in the past.

Production performances ARIMA vs ANN (Oct 2017 – Sep 2018)

Once in production, the model confirmed the simulation results by having an average error of ≈ 4% and leading the company to an economic benefit of ≈ 5.5M € in a year. The max MAPE was 24% that is 5% point over the simulated value but compared with the max MAPE achieved by ARIMA of 45%, it is much better. Both average and standard deviation of MAPE are 2% points better than the ARIMA model.

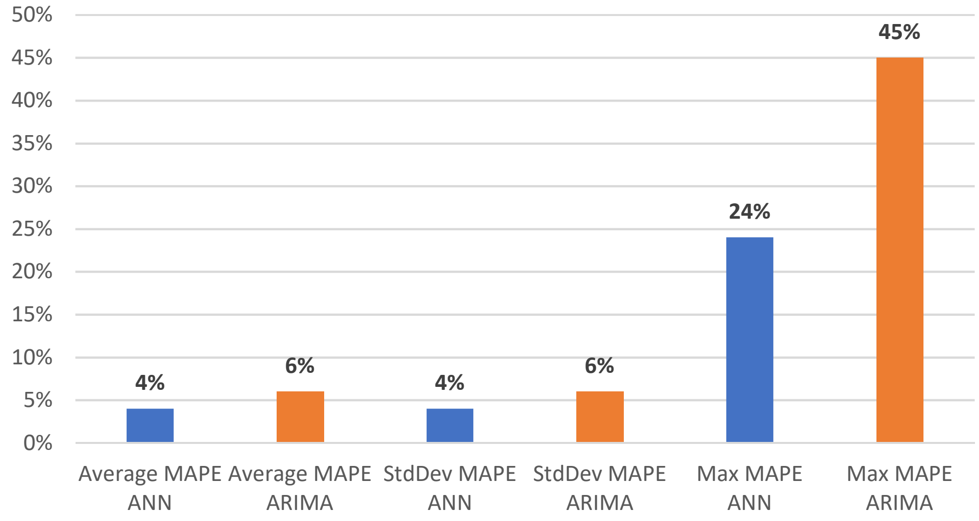


Figure 48 – MAPE comparison between ANN and ARIMA in production

During the implementation there was only one critical month in which the company lost incentives, in April 2017 it paid a fine of ≈ 2M€. There is another negative month August 2017 in which it paid fine. Nevertheless, the ANN model worked very well bringing more revenues to the company than what previously simulated. Instead, the old ARIMA model would have led the company to a big loss of ≈ 6.5M €. As we can see from real data the model ensures the performances calculated in advance during the simulation phase, having similar MAPE and similar MAPE standard deviation (≈ 4%).

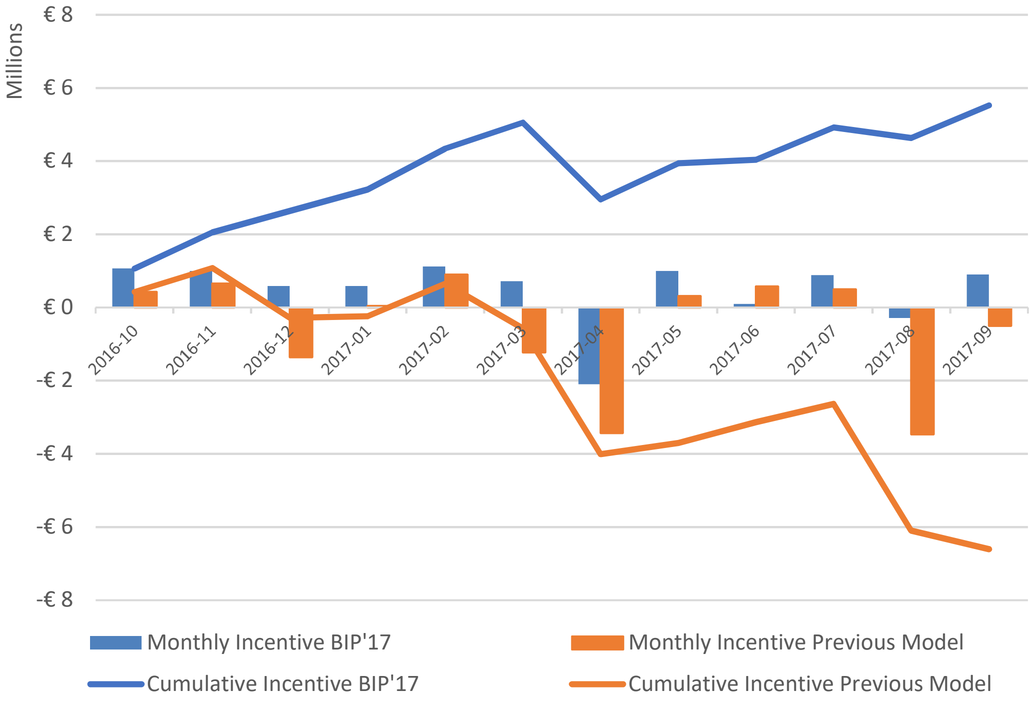


Figure 49 - Incentives comparison between ANN and ARIMA in production

This new model has an average error of ≈ 4%, approximately 1% less than the previous ARIMA model. But this new model had some limits, it worked well during the winter months, but in April 2018 unique gas demand patterns led to a monthly performance above 9%.

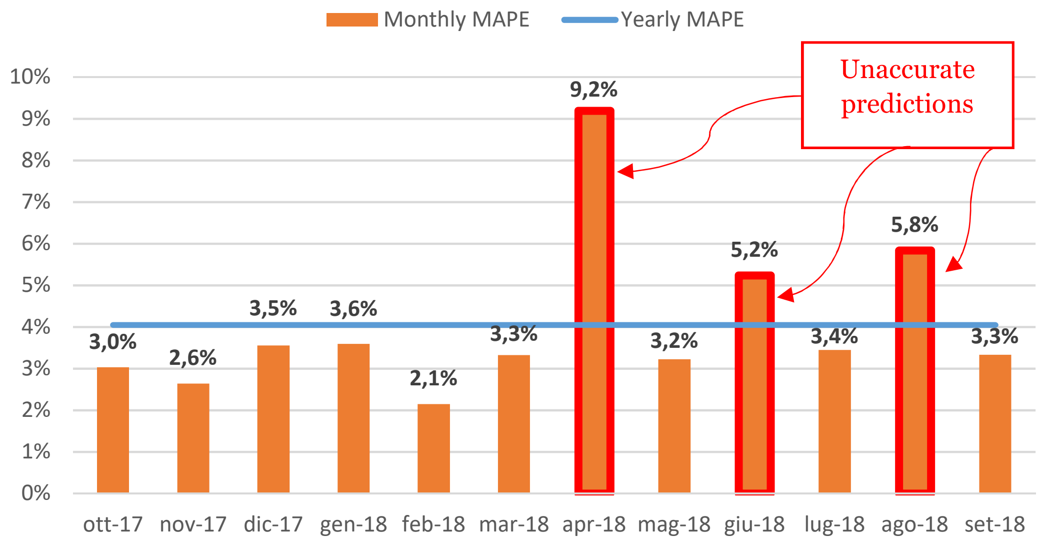


Figure 50 – Histogram figuring the MAPE for each month achieved by ANN, in blue the yearly MAPE

In red are remarked the critical months in which the average MAPE is over the yearly MAPE of ANN. The graph shows the main problem of the first neural network developed by Bip. During winter months it performs very well but in summer months not, and so starting from October 2018 Bip improved its model for the gas company in order to solve the summer mistakes in predictions.

4.5 Bip second solution: Ensemble

In October 2018 a new forecasting system was released, consisting of an ensemble of 6 sub-models characterized by different data sources and model architectures. The input data consists of the previously described sources as well as two new ones: electricity demand forecasts and more detailed weather forecasts.

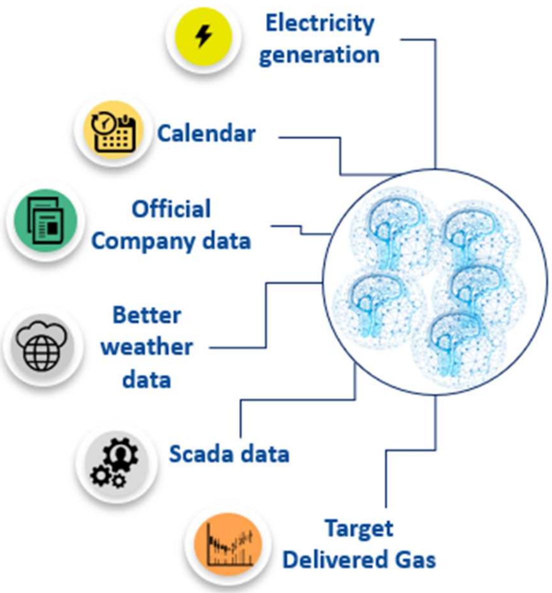


Figure 51 - Data input in Ensemble

The six sub-models are build using the following data:

Type of data	Description	Company Model	Scada Model	Complete Model	Electric Model	Similarity Model	Autoreg Model
Scada	Hourly data associated to entry and exit points	✓	✓	✓		✓	
Weather	Historical data and forecasts	✓	✓	✓	✓	✓	✓
Company	Official and provisional company data	✓		✓		✓	
Electricity	Provides knowledge about thermoelectric consumption	✓		✓	✓	✓	
Target	Historical target data, to capture the trend and seasonality	✓	✓	✓	✓		✓
Calendar	Binary variables for days etc.	✓		✓	✓	✓	✓

Table 18 – Table defining the data used to build all the six models

The prediction is based on each model weighted on their performances. Ensemble automatically make the predictions putting together all the six predictions of the sub-models:

- Every model contributes to the final prediction based on its performances
- The performance is calculated considering a dynamic time period

The main characteristics related to the Ensemble project are:

Ensemble
New functionalities: longer forecasting horizon, predictions per type of use
Duration: 28 weeks
3.5 FTE (Full Time Employee)
Prediction error below 3% (MAPE)
+30% improvement of the system w.r.t. baseline
+8M€ gained in 8 months
Year 2018

Table 19 – Ensemble main characteristics

In order to have a clear picture of the improvements led by Ensemble, the following are the simulation performances (Oct 2017 – Sep 2018) and the production performances (Oct 2018 – May 2019).

Simulation Performances ANN vs Ensemble (Oct 2017 – Sep 2018)

The simulation performances were built on the time period in which was operative the first ANN developed by Bip.

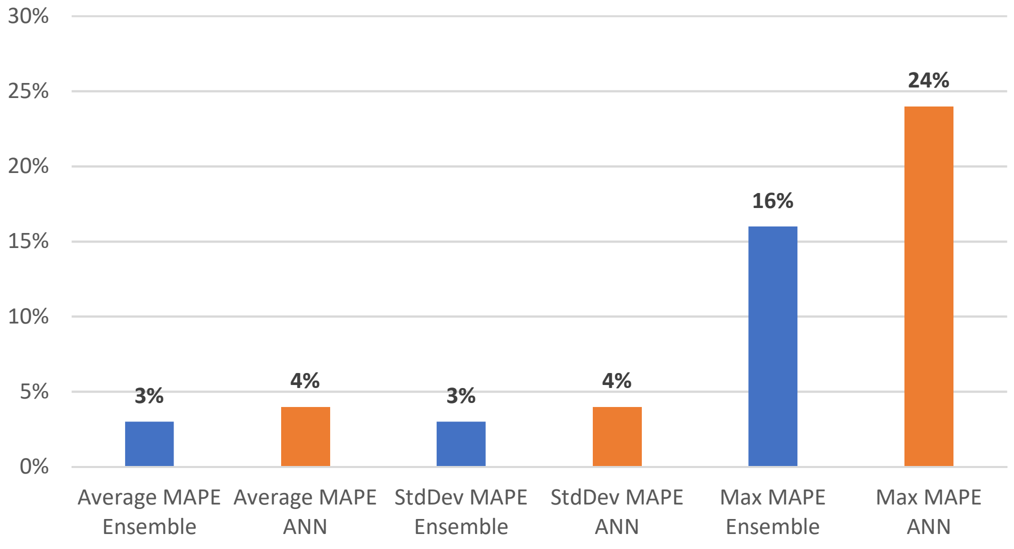


Figure 52 - MAPE comparison between Ensemble and ANN in simulation

The aim was to compare the real performances of the ANN with the new model Ensemble before to put it in production and the following are the results.

The simulation says that Ensemble reduces the average error of 1%, and even the variability is reduced a lot, since the maximum MAPE is 16%. Even the standard deviation of the MAPE is reduced by 1 % and the maximum MAPE obtained during the time period was 8% less than the maximum MAPE obtained by the ANN developed the year before. Having obtained these optimal results in the simulation, on Oct 1 Ensemble was released and put in production.

Production Performances ANN vs Ensemble (Oct 2018 – May 2019)

Data analyzed are available until May 2019 and not beyond, so the time period analyzed for the real case is not a year but eight months, precisely from October 2018 to May 2019. The following table will show how Ensemble performs:

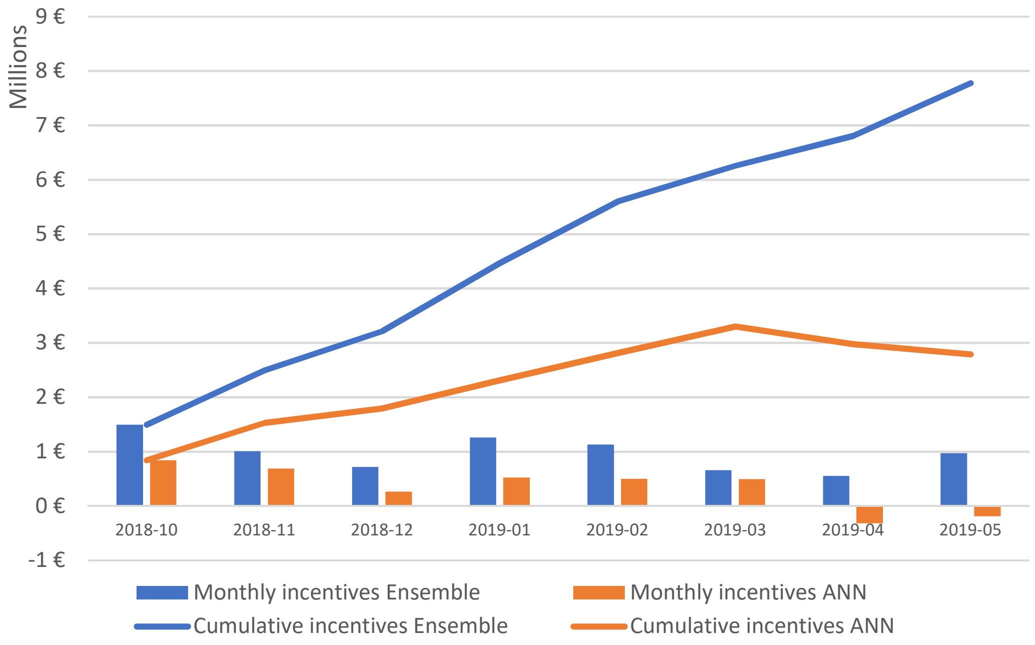


Figure 53 - Incentives comparison between Ensemble and ANN in production

Ensemble in eight months brought to the gas company a lot of incentives ≈ 8M €, instead of the precedent ANN would have brought just ≈ 3M €. Ensemble has achieved an enormous improvement under the economic point of view, transforming a threat imposed by the regulation law in a real strength for the company. Indeed, in April that was the biggest problem in prediction, with Ensemble we have a positive gain where instead ANN brings to an economic loss. Ensemble outperformed the previous ANN model because it reduced the average error and managed with care the summer period where the ANN had problems. Being an ensemble of models, it leverages on the predictions of 6 different models, weighting their results and providing accurate final predictions. It is a complicated neural network able to predict hourly the gas demand leveraging on many and very different types of data.

Ensemble gave a boost to the subsidies gained from the regulation law for the day-ahead predictions imposed by the authority, in fact, the deltas between all the different solution are the following:

	ARIMA	ANN	Ensemble	Δ
Incentives simulation Oct 2016 – Sep 2017 **[€]**	- 0.5M	+5M	\	+5.5M
Incentives production Oct 2017 – Sep 2018 **[€]**	-6.5M	+5.5M	\	+12M
Incentives production Oct 2018 – Sep 2019 **[€]**	\	+3M	+8M	+5M

Table 20 – Economic results between the three solutions in different periods of time

Every day there is a huge amount of money that could be gained or lost due to incentives for day-ahead forecasts. Gas company thanks to Bip has become data-driven and now it is able to deliver very accurate predictions for day-ahead. It is still working with Bip to maintain a continuous improvement of the performances.

Now that the results of the ML projects are clear we can discuss the objective of this fourth chapter: to create a comparison between performances obtained by the traditional ML models and the performances obtained by models created by the five different AutoML systems tested.

4.6 AutoML experiments

This section regards the experiments conducted on the gas company datasets to make a prediction for the day ahead. The datasets were appropriately anonymized by masking all the information that could lead back to the specific context.

The evaluation of each experiment is to measure the performances obtained by each AutoML tool used. The tools used were already presented in the market analysis (chapter 3.4) and in more detail with the benchmark regarding these systems (Chapter 3.6) and they are Google Cloud AutoML, Dataiku, H2O Driverless AI, AWS Sagemaker and Azure ML Microsoft.

Figure 54 - AutoML tools used for the business case, from left to right: Google AutoML Tables, Dataiku, Azure, AWS Sagemaker, and H2O Driverless AI

The objective is to compare the different performances obtained by the models created with these tools with the performances obtained by Ensemble and ANN. In this perspective before introducing the performances obtained from different tools, we need to explain the datasets used to train and to test the models. The different tools were tested on the same datasets, to build a consistent and meaningful comparison. The performances were collected for different phases of the pipeline, precisely for data cleaning feeding the tools with the raw dataset, for the feature engineering with the clean dataset, for the feature reduction with the engineered dataset, for the model selection with the reduced dataset and for the hyper-parameters optimization with the reduced dataset yet. Testing the tools in this way allows us to understand if a tool works better when a dataset has already some preprocessing steps or not. In the following table are reported the main characteristics of the datasets used.

Characteristics of Datasets

Dataset		Column	Row	Type of data	Size (Mb)
Raw	train	2,820	1,460	Date, numerical	40.3
	test	2,820	30	Date, numerical	0.907
Clean	train	2,820	1,460	Date, numerical	40.4
	test	2,820	30	Date, numerical	0.907
Feature engineered	train	2,919	1,460	Date, numerical	42.1
	test	2,919	30	Date, numerical	0.941
Feature reduced	train	125	1,460	Date, numerical	1.65
	test	125	30	Date, numerical	0.036

Table 21 – Characteristics of each dataset used in business case

The different models were tested on the test set representing October 2017, the following table define the time periods considered for the train set and the test set.

Target	***Tools***	**Train dataset**		**Test dataset**	
		From	***To***	***From***	***To***
October 2017	All	01/10/2013	30/09/2017	01/10/2017	31/10/2017

Table 22 – Three different target with tree different timeframe structures

The following picture is to build a visual representation of the training and the test set, the picture shows the time frame used to build models with the target of October 2017:

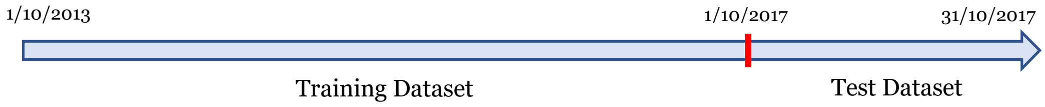


Figure 55 - Temporal split between training datasets and test datasets, considering 10/2017 as the test set

Then for each model was created a set of metrics to evaluate its performing state. The metrics used to evaluate the models are the following:

- MAPE (Mean Absolute Percentage Error)
  It is one of the most used metrics to assess the goodness of a regression model and it was already mentioned in Chapter 1.4.6 regarding the phase of the model evaluation. The reason why was chosen this metric is that it is directly related to the regulatory law and with the possibility to define precisely the economic return in adopting a certain model.

- Gain/Loss of the different solutions, comparing the AutoML models with the ANN and Ensemble models. As it was mentioned in the above point, the gain or loss amount of money due to the regulatory problem is directly related to the MAPE according to the Figure 39 which explains the relation between MAPE and incentives.
- Time, cost and human effort involved to build the different solutions, both ML and AutoML. These are all important drivers to take into consideration when the alternatives are under evaluation before starting a project.

Finally, it is time to analyze the different models built with the different software selected. The experiments analysis will start with Google Cloud AutoML Tables in the next paragraph.

4.6.1 Google Cloud AutoML Tables

The first tool experimented was AutoML Tables of Google.
The practical process followed to create the models is:

- Create an account
- Import training datasets
- Define the kind of problem (prediction)
- Define the target to predict
- Launch the training (1h default for every model trained)
- Get the model
- Import test datasets
- Evaluate the model on the test dataset (metric: MAPE)

These are the results for each model developed by Google Cloud AutoML:

	Data cleaning	*Features engineering*	*Features reduction*	*Model selection*	*Hyper-parameters optimization*	*ANN MAPE*	*Ensemble MAPE*
Dataset (10_2017)	Raw	Cleaned	Engineered	Reduced	/	3.07	1.56
Train error (%)	2.86	2.97	2.51	2.27	/		
Test error (%)	6.17	7.04	6.79	7.06	/		

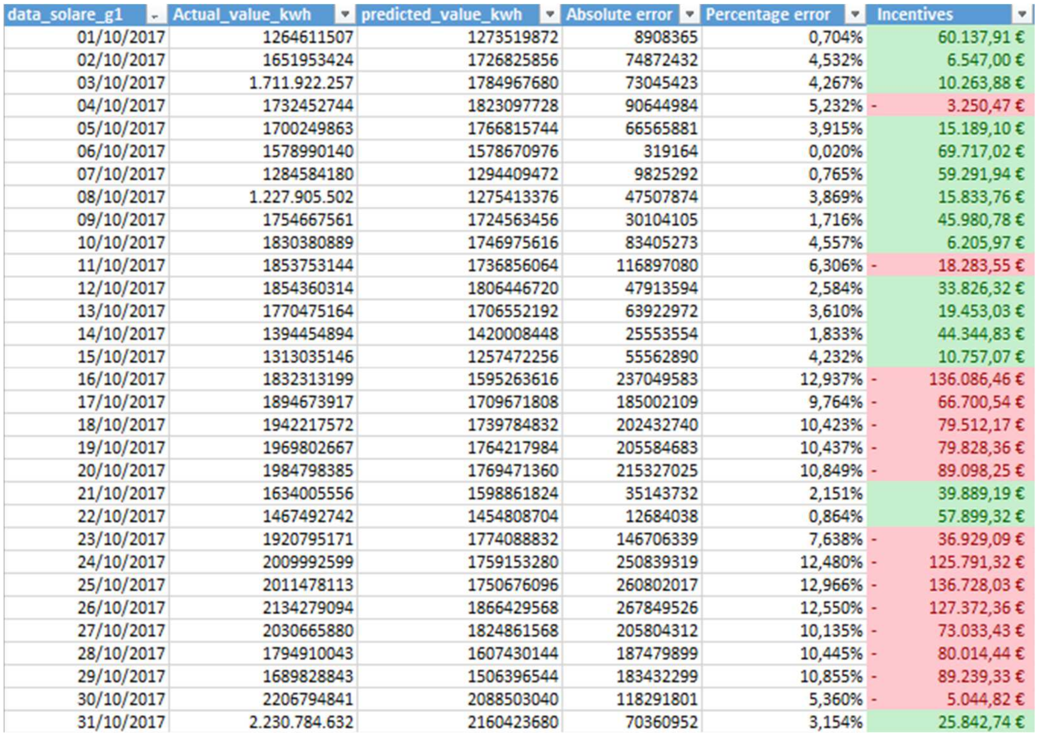
Table 23 – Performances obtained using Google AutoML Tables

The performances were calculated with the different situation as explained before. With Google was possible to create 4 different models based on 4 different configurations of the same training dataset (Table 19).

In this case, the best result in the test error is 6.17% when the dataset put in the software is in the raw condition. The result is not so good if compared with the reference performance of Ensemble. In this case, the bad performance is counterbalanced by the fact that AutoML Tables had achieved these results by running just for one hour to train a model. An essential thing to remember is that AutoML Tables validation was made considering as a test dataset October 2017 (winter month with a null incentive when MAPE = 5%). To give an economic point of view we can calculate what are the effective gains and losses during October 2017, calculating for each day the relative incentive and then summing them to have the total monthly incentives. This operation was made on Excel and the steps were:

- Calculate the relative error per each day considering the prediction and the real value.
- Calculate for each day the relative gain/loss applying a conditional formula describing how to calculate the incentive considering the relative error. Since October is in winter months the null incentive is matched when the percentage error is 5%.

Incentives = IF(Error<5% ; -(Error *100-5)*14000 ; IF(Error >10% ;
-(70000+(Error *100- 10)*22500) ; -(Error *100-5)*14000))



data_solare_g1	Actual_value_kwh	predicted_value_kwh	Absolute error	Percentage error	Incentives
01/10/2017	1264611507	1273519872	8908365	0,704%	60.137,91 €
02/10/2017	1651953424	1726825856	74872432	4,532%	6.547,00 €
03/10/2017	1.711.922.257	1784967680	73045423	4,267%	10.263,88 €
04/10/2017	1732452744	1823097728	90644984	5,232%	- 3.250,47 €
05/10/2017	1700249863	1766815744	66565881	3,915%	15.189,10 €
06/10/2017	1578990140	1578670976	319164	0,020%	69.717,02 €
07/10/2017	1284584180	1294409472	9825292	0,765%	59.291,94 €
08/10/2017	1.227.905.502	1275413376	47507874	3,869%	15.833,76 €
09/10/2017	1754667561	1724563456	30104105	1,716%	45.980,78 €
10/10/2017	1830380889	1746975616	83405273	4,557%	6.205,97 €
11/10/2017	1853753144	1736856064	116897080	6,306%	- 18.283,55 €
12/10/2017	1854360314	1806446720	47913594	2,584%	33.826,32 €
13/10/2017	1770475164	1706552192	63922972	3,610%	19.453,03 €
14/10/2017	1394454894	1420008448	25553554	1,833%	44.344,83 €
15/10/2017	1313035146	1257472256	55562890	4,232%	10.757,07 €
16/10/2017	1832313199	1595263616	237049583	12,937%	- 136.086,46 €
17/10/2017	1894673917	1709671808	185002109	9,764%	- 66.700,54 €
18/10/2017	1942217572	1739784832	202432740	10,423%	- 79.512,17 €
19/10/2017	1969802667	1764217984	205584683	10,437%	- 79.828,36 €
20/10/2017	1984798385	1769471360	215327025	10,849%	- 89.098,25 €
21/10/2017	1634005556	1598861824	35143732	2,151%	39.889,19 €
22/10/2017	1467492742	1454808704	12684038	0,864%	57.899,32 €
23/10/2017	1920795171	1774088832	146706339	7,638%	- 36.929,09 €
24/10/2017	2009992599	1759153280	250839319	12,480%	- 125.791,32 €
25/10/2017	2011478113	1750676096	260802017	12,966%	- 136.728,03 €
26/10/2017	2134279094	1866429568	267849526	12,550%	- 127.372,36 €
27/10/2017	2030665880	1824861568	205804312	10,135%	- 73.033,43 €
28/10/2017	1794910043	1607430144	187479899	10,445%	- 80.014,44 €
29/10/2017	1689828843	1506396544	183432299	10,855%	- 89.239,33 €
30/10/2017	2206794841	2088503040	118291801	5,360%	- 5.044,82 €
31/10/2017	2.230.784.632	2160423680	70360952	3,154%	25.842,74 €

Figure 56 – Excel page showing the steps to calculate the incentives gained or lost

The above picture shows for October 2017 the different steps to calculate incentives per each day. The table contains the predictions made by the model trained with raw data, it shows the relative error per day with the relative amount of incentive gained or lost during October 2017. Visually the incentives during this month are:

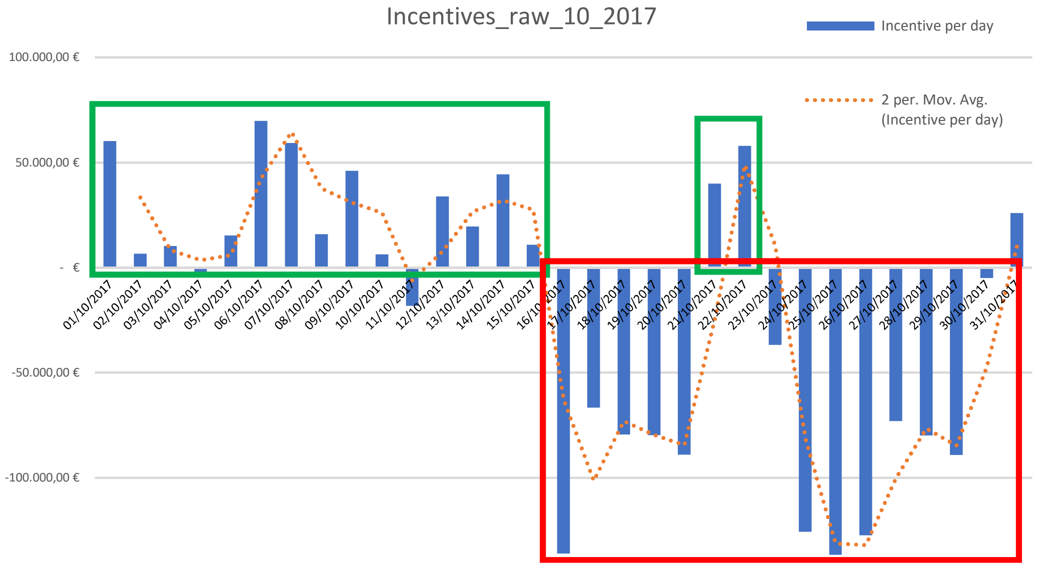


Figure 57 - Incentives during October 2017 considering the model built on the raw dataset (the best obtained by Google)

The chart highlights a critical situation from 16 October, indeed there are huge losses until 30/10. Instead in the first half of the month, there is a good situation in which the company benefits from the regulation. The following table put together all the monthly incentives applying the same steps to all the datasets tested with Google Cloud AutoML Tables, the performances and incentives refer only to the test period of October 2018.

It is not reported the step of hyper-parameters optimization because it was not executed by the tools as we said before.

	Raw	Cleaned	Engineered	Reduction
MAPE 10/2017 (%)	6.166	7.040	6.792	7.058
Tot incentives (€)	- 625,732	- 1,067,492	- 913,599	- 970,620

Table 24 – MAPE and incentives obtained by the different models built by GCP

From analysis emerged that the best case is obtained with raw dataset because the MAPE (6.166%) is the best and the economic losses are the smaller – 625,732.76 €, in the other case the MAPE is higher and the losses are ≈ 1M €.

4.6.2 Dataiku

Dataiku is the second AutoML system tested, the experiments followed these steps:

- Create a Dataiku account (free account)
- Import training datasets
- Define the problem to solve (prediction)
- Define the target (automatically detected)
- Launch the training phase (it lasts until the best model is not found)
- Select the best model
- Import test datasets, on which make predictions.
- Evaluate the model (metric: MAPE)

For Dataiku there are further experiments conducted to assess the goodness of the tool to create models, in particular were created and tested models for others two time periods, February 2018 and June 2018.

Target	**Train dataset**		**Test dataset**	
	from	***to***	***from***	***to***
February 2018	01/02/2014	31/01/2018	01/02/2018	28/02/2018
June 2018	01/06/2014	31/05/2018	01/06/2018	30/06/2018

Table 25 - Time periods of the successive tests conducted with Dataiku

In the following table there are the results for the different models built on training set for October 2017:

	Data cleaning	***Features engineering***	***Features reduction***	***Model selection***	***Hyper-parameters optimization***	***ANN MAPE***	***Ensemble MAPE***
Dataset (10_2017)	Raw	Cleaned	Engineered	Reduced	/	3.07	1.56
Train error (%)	0.58	0.77	0.44	0.84	/		
Test error (%)	5.81	11.35	5.47	4.53	/		

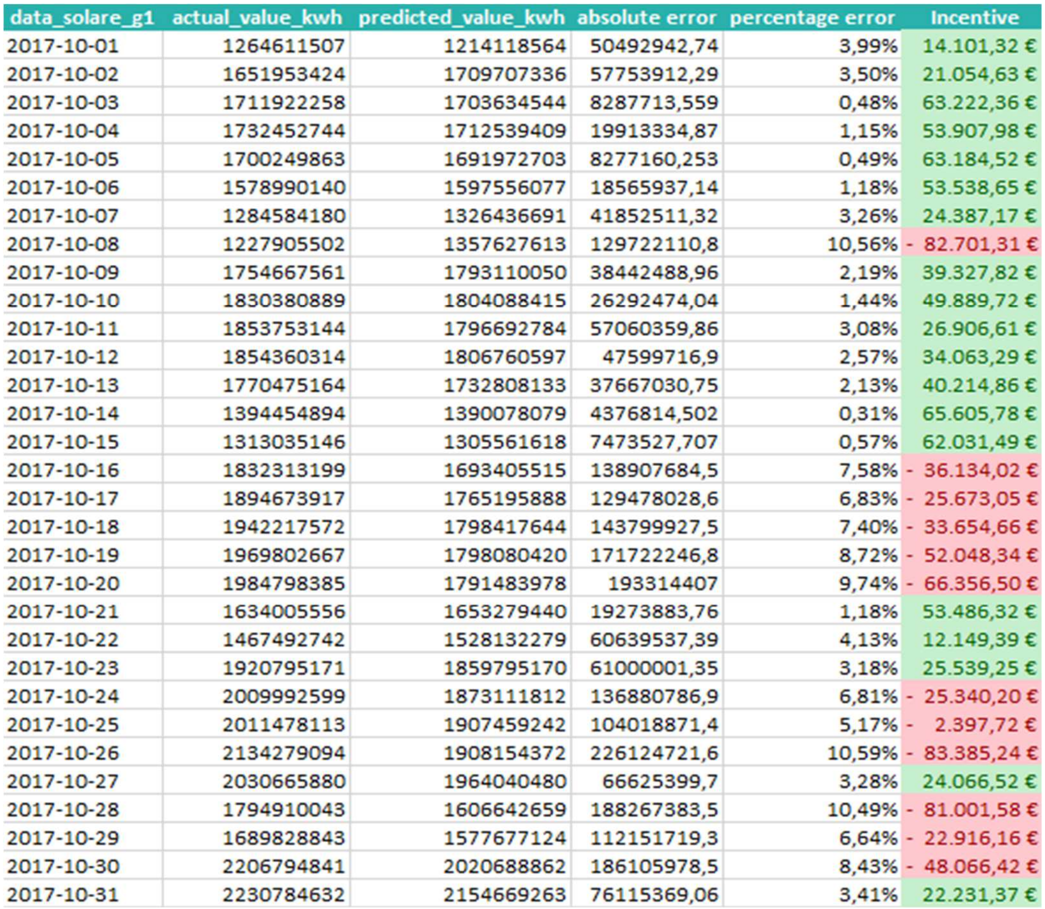
Table 26 – Performances obtained using Dataiku

October 2017

As we can see from performances table, in the train error the results are very good below 1% of MAPE, instead the best test error is achieved in model selection where the MAPE is 4.53%. The gap between the train and the test error is caused because probably Dataiku doesn’t manage well the overfitting problem, giving importance to features that in the train are relevant but that in dataset never seen before are not.

Nevertheless, the overall performances are quite good, the best result obtained from the prediction on October 2017 was 4.53% of MAPE and considering the threshold of the winter null incentive of 5% of MAPE, Dataiku achieves positive performances.

Taking in consideration the best model developed to predict the gas-demand for October 2017, precisely the model trained on the reduced dataset, the following figure is an Excel on which was calculated the relative incentives per each day. The relative incentives gained potentially by the company using the Dataiku best model during October 2017 are:



data_solare_g1	actual_value_kwh	predicted_value_kwh	absolute error	percentage error	Incentive
2017-10-01	1264611507	1214118564	50492942,74	3,99%	14.101,32 €
2017-10-02	1651953424	1709707336	57753912,29	3,50%	21.054,63 €
2017-10-03	1711922258	1703634544	8287713,559	0,48%	63.222,36 €
2017-10-04	1732452744	1712539409	19913334,87	1,15%	53.907,98 €
2017-10-05	1700249863	1691972703	8277160,253	0,49%	63.184,52 €
2017-10-06	1578990140	1597556077	18565937,14	1,18%	53.538,65 €
2017-10-07	1284584180	1326436691	41852511,32	3,26%	24.387,17 €
2017-10-08	1227905502	1357627613	129722110,8	10,56%	- 82.701,31 €
2017-10-09	1754667561	1793110050	38442488,96	2,19%	39.327,82 €
2017-10-10	1830380889	1804088415	26292474,04	1,44%	49.889,72 €
2017-10-11	1853753144	1796692784	57060359,86	3,08%	26.906,61 €
2017-10-12	1854360314	1806760597	47599716,9	2,57%	34.063,29 €
2017-10-13	1770475164	1732808133	37667030,75	2,13%	40.214,86 €
2017-10-14	1394454894	1390078079	4376814,502	0,31%	65.605,78 €
2017-10-15	1313035146	1305561618	7473527,707	0,57%	62.031,49 €
2017-10-16	1832313199	1693405515	138907684,5	7,58%	- 36.134,02 €
2017-10-17	1894673917	1765195888	129478028,6	6,83%	- 25.673,05 €
2017-10-18	1942217572	1798417644	143799927,5	7,40%	- 33.654,66 €
2017-10-19	1969802667	1798080420	171722246,8	8,72%	- 52.048,34 €
2017-10-20	1984798385	1791483978	193314407	9,74%	- 66.356,50 €
2017-10-21	1634005556	1653279440	19273883,76	1,18%	53.486,32 €
2017-10-22	1467492742	1528132279	60639537,39	4,13%	12.149,39 €
2017-10-23	1920795171	1859795170	61000001,35	3,18%	25.539,25 €
2017-10-24	2009992599	1873111812	136880786,9	6,81%	- 25.340,20 €
2017-10-25	2011478113	1907459242	104018871,4	5,17%	- 2.397,72 €
2017-10-26	2134279094	1908154372	226124721,6	10,59%	- 83.385,24 €
2017-10-27	2030665880	1964040480	66625399,7	3,28%	24.066,52 €
2017-10-28	1794910043	1606642659	188267383,5	10,49%	- 81.001,58 €
2017-10-29	1689828843	1577677124	112151719,3	6,64%	- 22.916,16 €
2017-10-30	2206794841	2020688862	186105978,5	8,43%	- 48.066,42 €
2017-10-31	2230784632	2154669263	76115369,06	3,41%	22.231,37 €

Figure 58 – Excel page showing the steps to calculate the incentives gained or lost

Testing Dataiku automated ML model bring us to have an average MAPE = 4.53% that enables the company to respect the null incentive of 5%.

From the excel is already evident the tendency to have positive results in the first half of the month and bad results in the second part. The same happened with Google AutoML Tables, both the tools work well in the first half and work bad in the second. Visually the incentives during October 2017 using the model build by Dataiku are in the following chart:

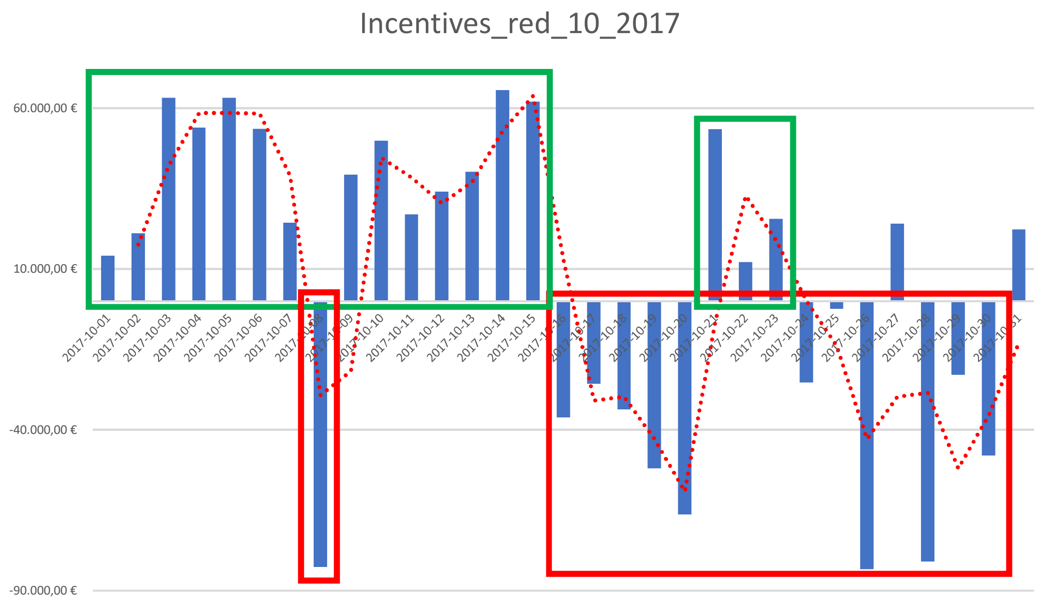


Figure 59 - Incentives during October 2017 considering the best model with MAPE of 4,53%

What emerged from the chart is that there is a specific day (2017-10-8) in the first half of the month that have a very bad performance, making lose to the gas company more than 80,000€. As in Google AutoML Tables experiments, bad performances began from 2017-10-16 and last for the rest of the month, with few positive exceptions.

The overall incentives considering the different models built using Dataiku are:

		Raw	**Cleaned**	**Engineered**	**Reduced**
October 2017	**MAPE (%)**	5.81	11.35	5.47	4.53
	Incentives (€)	- 397,886	- 3,940,372	- 268,311	189,233

Table 27 – MAPE and incentives obtained by the different models built by Dataiku

As we said before with Dataiku were testes others models to predict others two different target periods. The performances obtained by these tests are in the following tables:

	Data cleaning	***Features engineering***	***Features reduction***	***Model selection***	***Hyper-parameters optimization***
Dataset (2_2018)	Raw	Cleaned	Engineered	Reduced	/
Train error (%)	3.38	3.39	3.41	2.97	/
Test error (%)	4.05	3.65	3.35	2.79	/
Dataset (6_2018)	Raw	Cleaned	Engineered	Reduced	/
Train error (%)	3.64	3.71	3.44	3.35	/
Test error (%)	2.30	2.25	24.59	2.80	/

Table 28 - Performances obtained by Dataiku on further tests made on 2/2018 and 6/2018

What emerged is that:

- February 2018
  The performances obtained in train test and test set are similar, and they are all good considering the null incentive of 5% of the winter period. The best MAPE is obtained by the model fed by the reduced train dataset with a MAPE equal to 2.79%. During these experiments there are no critical situations to highlight.
- June 2018
  Even for June the predictions made by the different model developed are good, achieving the best result with the model created from the raw dataset with a relative MAPE of 2.30%. There is a critical situation to highlight, the MAPE obtained by the test on the engineered dataset has a MAPE of 24.59%, a very bad result.

As was made for October 2017, even for the other two target periods is useful understand how many incentives the different models created could gain or lose.

		Raw	**Cleaned**	**Engineered**	**Reduced**
February 2018	**MAPE (%)**	4.05	3.65	3.35	2.79
	Incentives (€)	348,641	485,639	601,886	868,103
June 2018	**MAPE (%)**	2.30	2.25	24.59	2.80
	Incentives (€)	1,132,697	1,135,716	- 12,046,320	981,168

Table 29 - MAPE and incentives obtained by the different models built by Dataiku on further target periods

From the experiments conducted on Dataiku, considering also the models created to predict other time periods, the insights are:

- For each months the performances achieve good results, for each month there is at least on model that respect the threshold of 5%.
- There are two critical situations to highlight. The first was during the test on October 2017 when was achieved a MAPE equal to 11.35% and the second was during the test on June 2018 when was achieved a MAPE equal to 24.59%.
- For October 2017 only one model respects the null incentive threshold of 5%, it is the model trained with the reduced dataset.
- The best results are obtained by the prediction on June 2018, where the ANN was over the threshold of 5%.
- From an economic point of view, different models of Dataiku allow to gain positive incentives if they would have been adopted by the gas company.
- Dataiku is very fast to train model: considering the model development with the reduced dataset, the training model is between 5-15 minutes both for Neural Network and XGBoost algorithms.

4.6.3 Azure ML Microsoft

The third tool of AutoML tested is Azure Machine Learning, the following are the practical steps of the experimentation:

- Create an Azure account
- Create an Azure ML workspace
- Import training datasets
- Definition of the problem to solve (prediction)
- Launch the training phase
- Get the best model
- Import test datasets
- Evaluate the model on test datasets (metric: MAPE)

On the website of Azure is available a figure that explain how their AutoML services work. Azure request datasets in input on which train models, it requests the target is manually defined by the operator and in the end, it requests the input of constraints limits as running time and cost of machines used. Then it develops and evaluates different models, presented to the users on a leaderboard with the relative score of validation.

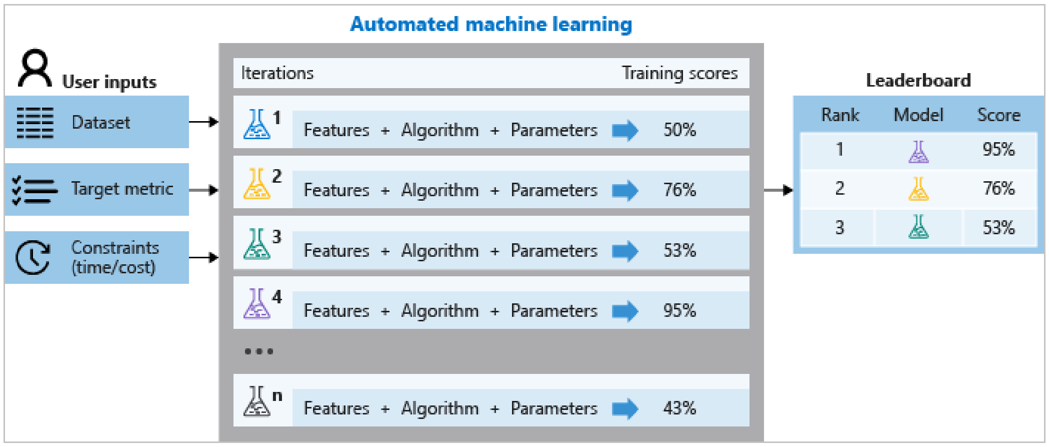

Automated machine learning
User inputs
Dataset
Target metric
Constraints (time/cost)
Iterations
Training scores
1 Features + Algorithm + Parameters 50%
2 Features + Algorithm + Parameters 76%
3 Features + Algorithm + Parameters 53%
4 Features + Algorithm + Parameters 95%
...
n Features + Algorithm + Parameters 43%
Leaderboard
Rank
Model
Score
1 95%
2 76%
3 53%

Figure 60 – Azure workflow

Considering our case, the relative metrics obtained for the different model created are in the following table:

	Data cleaning	*Features engineering*	*Features reduction*	*Model selection*	*Hyper-parameters optimization*	*ANN MAPE (%)*	*Ensemble MAPE (%)*
Dataset (10_2017)	Raw	Cleaned	/	Reduced	Reduced		
Train error (%)	2.70	2.70	/	2.60	2.60	3.07	1.56
Test error (%)	5.60	5.30	/	3.70	3.60		

Table 30 - Performances obtained using Azure

Azure doesn't perform the feature reduction phase, need human intervention. The test error is 3.6% that is quite good but not as good as the train error equal to 2.6%. There is a consistent improvement in model selection, achieving the best result when the tool deals with hyper-parameters optimization. Considering the four models created only the half are below the threshold of 5%. There are no performances better than Ensemble and ANN.

The average MAPE obtained by Azure Microsoft is better than Google AutoML Tables and Dataiku, but it is not as good as ANN and Ensemble. Azure can be a good solution to implement for the company because it achieves a MAPE lower than 5% allowing it to gain incentives, indeed considering October 2017 Azure would have brought ≈ 490,000 €.

In the case of Azure the results are better than Google and Dataiku, with the following table we can compare it with all the previous solutions:

	Raw	**Cleaned**	**Engineered**	**Reduced**
MAPE 10/2017 (%)	5.6	5.3	/	3.6
Tot incentives (€)	- 377,580	- 247,380	/	490,420

Table 31 - MAPE and incentives obtained by the different models built by Azure

According with our business problem Azure achieves good results when the dataset is already cleaned, engineered and reduced. Compared with Google, it provides positive incentives and more accurate models. Comparing Azure and Dataiku there is an evident fact: the best results were obtained when the AutoML systems were ingested with an already cleaned and reduced datasets, they seem to work better if datasets are previously preprocessed.

4.6.4 H2O Driverless AI

The fourth AutoML software used is H2O Driverless AI.
It is a complete tool able to deal with all phases of ML pipeline. The practical steps were:

- Create an H2O Driverless AI environment
- Import training datasets
- Definition of the problem to solve (prediction)
- Launch the training phase
- Get the best model
- Import test datasets
- Evaluate the model on test datasets (metric: MAPE)

The different performances obtained by H2O Driverless AI are:

	Data cleaning	*Features engineering*	*Features reduction*	*Model selection*	*Hyper-parameters optimization*	*ANN MAPE (%)*	*Ensemble MAPE (%)*
Dataset (10_2017)	Raw	Cleaned	Engineered	Reduced	Reduced	3.07	1.56
Train error (%)	3.55	3.48	4.41	4.17	4.17		
Test error (%)	4.54	5.01	5.28	2.99	12.53		

Table 32 - Performances obtained using H2O Driverless AI

The best performance with the training dataset was achieved in the feature engineering phase with a MAPE of 3.48%, a good result compared with the ANN reference and a bad result if compared with Ensemble performance. Instead the performances obtained with the test dataset were very good in the phase of model selection achieving a MAPE below the 3%. From the forecasting model developed by H2O Driverless AI is evident that it is a good system to deal this kind of problems like prediction and forecasting.

There is an important crucial situation to highlight, it occurs in automating the hyper-parameters optimization. In this specific case the tool performed very bad, achieving the highest monthly MAPE considering all the tools tested until now.

Considering that H2O Driverless AI achieved with two models a monthly MAPE for October 2017 lower than 5%, we have two cases in which the incentives are positives. The following table shows the relative incentives gained considering every model built:

	Raw	**Cleaned**	**Engineered**	**Reduction**	**Hyper**
MAPE (%)	4.54	5.01	5.28	2.99	12.53
Tot incentives (€)	82,460	- 121,520	- 238,700	755,160	- 3,939,675

Table 33 - MAPE and incentives obtained by the different models built by H2O Driverless AI

4.6.5 AWS Sagemaker

In case of AWS Sagemaker it was tested because it is considered a very powerful tool created by a leader company. It was tested even if it automates only the hyper-parameter optimization. The steps of the experiment were:

- Create an AWS Sagemaker account (free)
- Import training datasets
- Definition of the problem to solve (prediction)
- Definition of the model to use
- Launch the hyper-parameters optimization
- Get the best model's parameters setting
- Import test datasets
- Evaluate the model on test datasets (metric: MAPE)

In the following table there is the only result coming from the experiment of AWS Sagemaker:

	Data cleaning	*Features engineering*	*Features reduction*	*Model selection*	*Hyper-parameters optimization*	*ANN MAPE (%)*	*Ensemble MAPE (%)*
Dataset (10_2017)	raw	cleaned	engineered	reduced	reduced	3.07	1.56
Train error (%)	/	/	/	/	1.39		
Test error (%)	/	/	/	/	5.17		

Table 34 - Performances obtained using AWS Sagemaker

In this case AWS Sagemaker has the only objective to improve the model finding the best combination of hyper-parameters. The system runs until it achieves the best hyper-parameters that allow the model to fit well the data. The best result achieved during the hyper-parameters optimization is 5.20% of MAPE. It is not a good result because it is more than 5% (null incentive) and it leads the company to have a loss during October 2017.

	Hyper-parameters Optimization
MAPE (%)	5.17
Tot incentives (€)	-203,980

Table 35 - MAPE and incentives obtained by the model developed by AWS Sagemaker

Adopting AWS to create the best setting of hyper-parameters automatically would lead the company to a consistent loss of -203.980 €. Considering that during October 2017 the average MAPE of the traditional ML projects are 3,07% for ANN and 1,56% for Ensemble we can conclude that AWS in solving this problem does return an accurate model to allow the gas company to make the regulation a strength for its business. Further tests are necessary to assess the potentialities of this tools in dealing with regression problems.

4.7 Final insights

We tested five different AutoML systems to build models with the aim to solve a very difficult prediction task, predicting the demand for the daily-ahead gas consumption. This problem is now solved by a ML model very complex that needed two years to achieve a yearly MAPE of 3%, we are talking about Ensemble model. In the following table we resume all the main important data to describe the performances of AutoML systems regarding the prediction task for October 2017:

	Google AutoML	***Azure Microsoft***	***Dataiku***	***H2O Driverless AI***	***AWS Sagemaker***	***ANN***	***Ensemble***
Best MAPE obtained 10/2017 (%)	6.17	3.60	4.53	2.99	5.20	3.07	1.56
Worst MAPE obtained 10/2017 (%)	7.06	5.60	5.50	12.53	5.20		
incentives best scenario (€)	-625,732	490,420	189,233	755,160	-203,980	724,780	1,380,120
incentives worst scenario (€)	-1,067,492	-377,580	-3,940,372	- 3,939,675	-203,980		
Project time (h)	4	4	8	4	4	960	1120
FTE	1	1	1	1	1	2.5	3.5
Internal costs (€)	125	125	250	125	125	75,000	122,500

Table 36 - Full comparison between AutoML solutions and ML solutions (ANN and Ensemble)

AutoML does not perform as well as traditional ML models in our business case, but in doing the evaluation of these systems we have also to consider the trade-off between the performances achieved by the models and time needed to develop models with relative costs. The main insights we can extract form Table 35 are:

- With 3 different tools were achieved positive performances, with H2O Driverless AI was even outperformed the MAPE obtained by ANN during October 2017.
- The time to develop models are drastically reduced compared to Ensemble and ANN models.
- The internal cost for Bip to develop model is directly affected by the number of data scientists dedicated and the time needed to develop the model. With AutoML the costs are heavily reduced.

The final evaluation of our AutoML tests could be based on the trade-off composed by: Performances, Incentives, Project Costs.

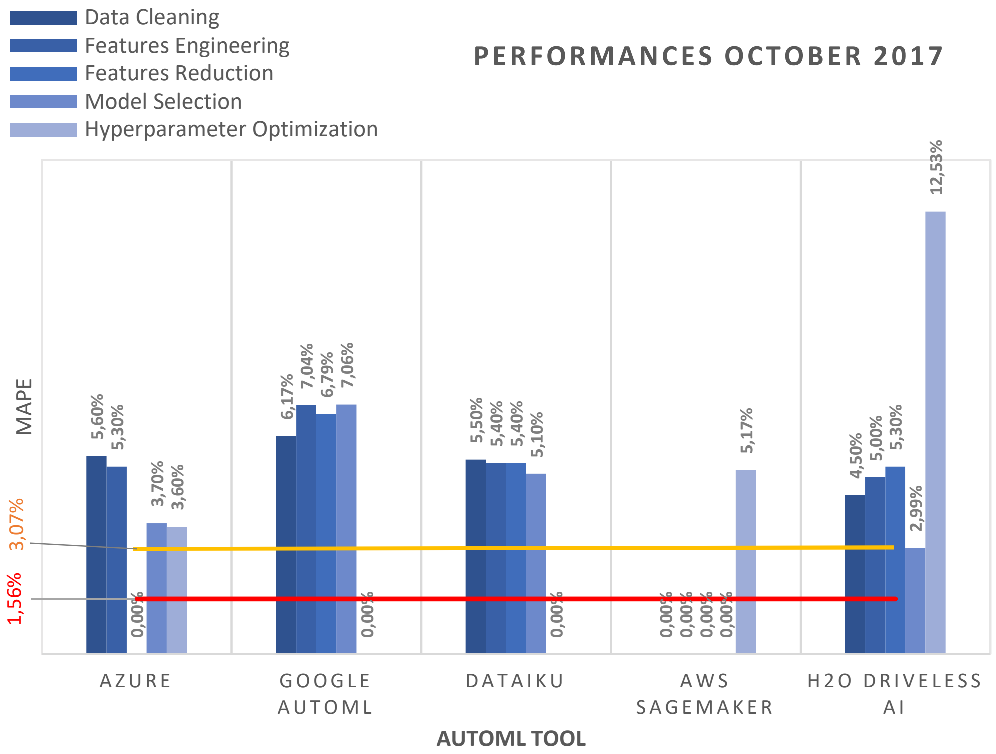


Figure 61 - Overall performances obtained by each tool for every model developed, comparing MAPE of every model with the two thresholds given by the monthly MAPE of ANN and by the monthly MAPE of Ensemble

To sum up all the performances obtained on the test dataset it was built the following histogram chart that explain for each AutoML tool the performances achieved by each model considering as a test dataset October 2017.

Figure 57 shows graphically the different performances obtained by the different models developed by each AutoML tool. It is evident that in average the models do not perform as well as ANN and Ensemble, mainly with Ensemble threshold no one model achieves its monthly performance of 1.56%. There is only one model developed with H2O Driverless AI that achieves and outperforms the threshold of ANN.

Figure 58 shows a comparison of incentives gained or lost, taking into consideration the best models developed by each tool. The chosen models are:

- Google cloud AutoML -> Data Cleaning, MAPE = 6.17%
- Azure -> Hyper-parameter optimization, MAPE = 3.60%
- Dataiku -> Model selection, MAPE = 4.53%
- H2O Driverless AI -> Model selection, MAPE = 2.99%
- AWS Sagemaker -> Hyper-parameters optimization, MAPE = 5.17%

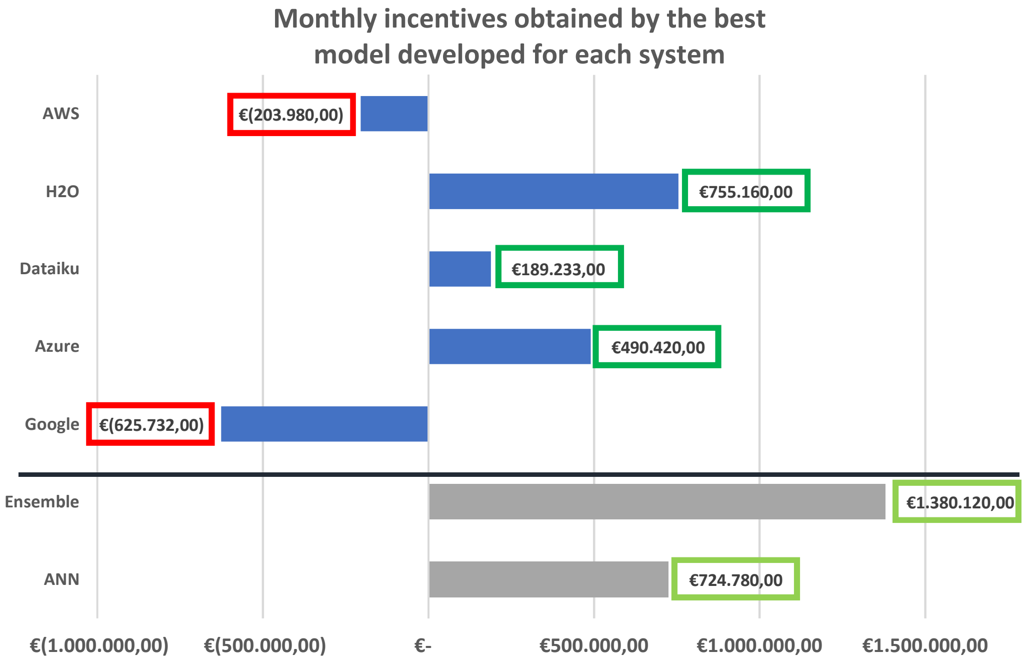


Figure 62 - Comparison between incentives gained or payed adopting AutoML or ML in October 2017

The above figure shows the comparison between the incentives that each model, if put in production by the gas company, would have brought to it during October 2017. We can see that both ANN and Ensemble bring positive incentives and they are the references of comparison for the incentives potentially gained or lost by the AutoML models. In 3 case out of 5 the tests make positive incentives, to highlight the amount of incentives gained by the model of h2O Driverless AI which achieves incentives higher than the incentives brought by ANN model. The worst predictive model is the one created by Google Cloud AutoML that potentially make lose a big amount of money to the company. The last topic of the trade-off analysis to explain is the amount of effort and cost needed to develop the different models. After having explained this perspective we will list the final insights about our tests for this kind of business problem. The next table will resume the number of persons dedicated to each model development with the relative cost and time needed to deploy the model and to make the predictions. Even this part is compared with ANN and Ensemble solutions.

System	Time [h]	FTE [p]	Cost [€]
ANN	960	2.5	75,000
Ensemble	1,120	3.5	122,500
Google	4	1	125
Azure	≈ 4	1	≈ 125
Dataiku	≈ 8	1	≈ 250
H2O Driverless AI	≈ 4	1	≈ 125
AWS Sagemaker	≈ 4	1	≈ 125

Table 37 - Comparison of 3 drivers: Time, Full Time Employee and Cost related to each solution

Where:

- **Time**: it is the whole time needed to train all the models for every system used.
- **FTE**: it stands for Full Time Employee, it means the number of persons allocated to a specific project.
- **Cost**: it is the internal cost for Bip in developing the different solution considering the person allocated and the amount of time needed. The internal daily cost for a person allocated to a project is 250 €.

It is evident the difference between the amount of effort, time and cost needed to develop the ANN and Ensemble model and the AutoML solutions. AutoML brings benefits under the point of view of time and costs related to the project, in our case considering Ensemble the time of the project was 28 weeks with 3.5 FTE, instead considering AutoML solutions there are no projects that run over 1 day allocating only 1 FTE.

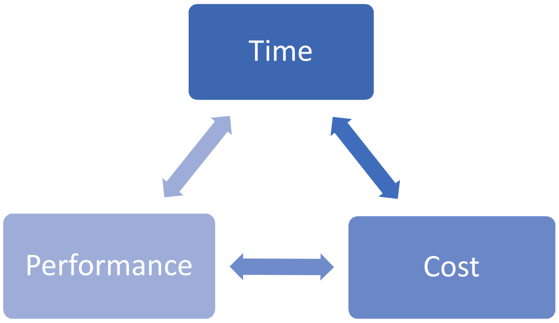


Figure 63 - Trade-off

From the comparison between our AutoML models and the ML models, considering the business context of the gas company considered, what emerged is that:

	Benefits	Constraints
ML	• High performances • High incentives	• Projects long at least 24 weeks • High internal costs • High efforts • High complexity
AutoML	• Very fast model development, even just 1 hour to train a model • Low internal costs • Some models achieve good performances under the null incentive	• No or very reduced possibility to customize the models • Some models do not achieve good performances • Some tools make potentially pay fines due to bad predictions • Some tools are black-box, you can't know how they build models.

Table 38 - Benefits and constraints emerged from the tests conducted on AutoML with our specific problem

5 CONCLUSIONS

This is the final chapter of the thesis, here we are going to discuss the insights extracted from all the analysis and tests made about the world of automated machine learning. The main purpose of this chapter is to understand if we satisfied the objectives defined at the beginning, and then to resume the main results. For this reason, we are going to construct this chapter following the 4 main objectives fixed in Chapter 1. For every objective will be dedicated to a paragraph, as follow.

To create knowledge and awareness

This is the first objective thesis tries to cover, it is focused on the reason why AutoML arose in the last years and what are the main benefits and constraints it tries to solve. This objective is satisfied in the whole thesis, mainly in Chapter 1 where were introduced the reasons why these new systems were born, and the main problems traditional ML is facing. The thesis explains the objectives of AutoML and how it works considering as a reference the traditional pipeline development, using in detail the CRISP-DM model to understand the utility of AutoML in each phase.

Concisely, what emerged during the thesis that satisfies this objective is:

- ✓ What is automated ML
- ✓ AutoML benefits and constraints
- ✓ Differences with traditional ML
- ✓ AutoML market, main solutions available today
- ✓ Organizational impact when adopting AutoML
- ✓ Performances obtained in a real business case with a prediction problem

To create the big picture of the actual tools

The second main objective defined at the beginning was to define what are the main available solutions in existence today, this was accomplished mapping the market and understanding in which industries AutoML is implemented as a real source of value to run businesses. This objective is satisfied by Chapter 3 in which we created a market analysis considering a total of 53 different AutoML systems that we analyzed under different points of view:

- ✓ Definition the type of tools.
- ✓ Definition of the capabilities of each tool.

- ✓ Market growth pace.
- ✓ Mapping the main customers of AutoML tools.
- ✓ Analysis of the customers per industry.
- ✓ Benchmark on 5 tools used in the experiments.

Organizational impact of AutoML

The third objective was to study the impact AutoML has on data science team organizations, considering the professional role of the data scientist, and the emerging professional role of the citizen data scientist emerged thanks to analytical tools like AutoML. The main results of this chapter are:

- ✓ Organizational impact considering the CRISP-DM Model and the traditional composition of a data science team.
- ✓ Impact on the role of data scientists, scenario analysis.
- ✓ Definition of the new emerging role of the citizen data scientist, analysis of 3 main articles published by Forbes, Forrester, and Gartner.

Comparison between ML and AutoML performances

To satisfy this need we chose a set of AutoML tools to test. The problem to solve is a real case developed in Bip.xTech, regarding a prediction problem for a gas company. The comparison was made considering as a metric the mean absolute percentage error (MAPE) calculated on a specific time period (October 2017), in order to compare the AutoML performances with the ANN and Ensemble models. Moreover, the comparison takes into consideration even the economic impact of AutoML comparing the potential incentives gained and internal costs for the different machine learning models developed.

The main insights extracted by AutoML tools during experiments are:

- ✓ AutoML does not perform as well as a customized ML model made by a team of data scientists.
- ✓ AutoML models achieve in average ‘good’ performances, where ‘good’ means relatively close performances to traditional ML ones. In one case the performance of ANN is beaten by the H2O Driverless AI one.
- ✓ AutoML models are cheaper, faster and easier to develop compared with traditional ML models.

The 4 main objectives of the thesis were satisfied, the only doubt is coming from the qualitative analysis of the impact this new technology will bring in the data science world and more specifically on data scientists. As we said in Chapter 2 the scenario most probable to become real is the one in which AutoML becomes a powerful support tool for data scientists improving their productivity and an empowering tool for citizen data scientists that will become an important figure inside companies. To consolidate this scenario a future work should be collect responses from surveys sent to data scientists in order to comprehend the overall thoughts about AutoML, if they use it, in which quantities, which tools, which part they automate and so on.

From the analysis made, we can assume that it is a growing market, since form the last 5 years we have seen an increment of AutoML tools of +269% (Chapter 3.4). These tools have a strong demand in industries where data are becoming more and more relevant over time, specifically in financial services, software, and healthcare.

As we have seen from the experiments the limits for AutoML solutions are the ability to understand the business context that must be set by a human intervention yet. Further, for our business problem they do not achieve as good performances as ML models, and sometimes they do not respect neither the null incentive threshold of 5% of MAPE. The confirmed benefits of AutoML are that several tools are very user friendly and they enable data scientists to accelerate the workflow of a ML project. Furthermore, they allow to avoid repetitive tasks like feature cleaning, feature engineering, and hyper-parameters optimization that are time-consuming for data scientists.

Could AutoML be considered as a valid alternative to traditional machine learning workflow?

This was the challenging question of the abstract, now we can answer using researches conducted and results obtained.

Yes. It is a very powerful technology, it creates value for the company saving time, costs and human effort. Since there are many AutoML tools with different characteristics, companies can choose the most useful to solve their problems. What we learn from the tests on the gas company is that this new technology is essential if there are narrow constraints in time and cost. In this case AutoML is timely and cheap, but if there are constraints on performances it does not perform as well as ML models developed by data scientists. The choice to adopt or not AutoML depends on the business constraints of time, budget and human resources allocated.

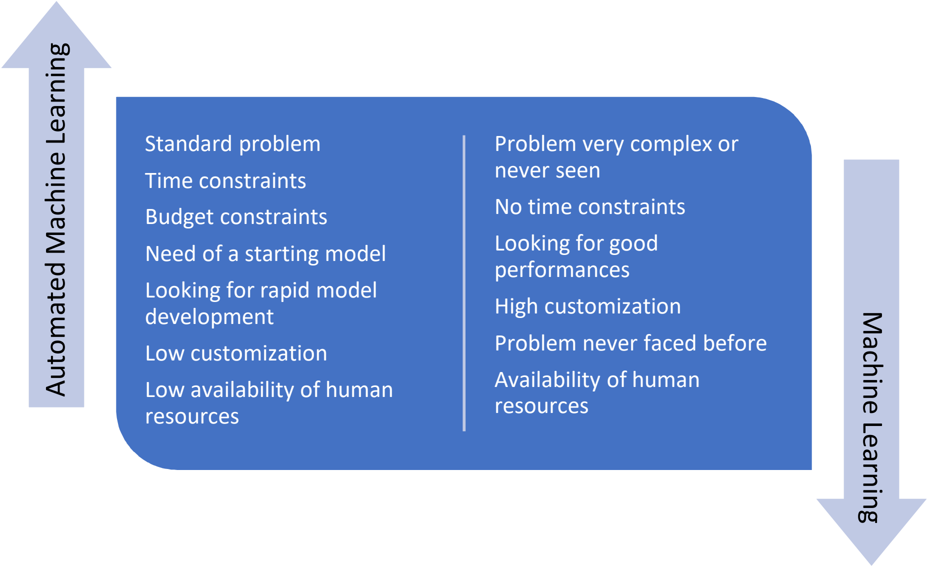

Automated Machine Learning
Standard problem
Time constraints
Budget constraints
Need of a starting model
Looking for rapid model development
Low customization
Low availability of human resources
Problem very complex or never seen
No time constraints
Looking for good performances
High customization
Problem never faced before
Availability of human resources
Machine Learning

Figure 64 - Drivers of the choice between AutoML and ML

5.1 Future researches and applications

This thesis explains at a high level which impact AutoML brings in the world of ML, defining the general market trends, defining who are the target customers of providers and why it is used by companies to improve their business. Then, we focused our attention on comparing the performances obtained by different AutoML tools on a prediction problem, already solved by the ML models Ensemble, to understand if these new solutions could replace the manual development of a ML model with this kind of problem. The average results of our experiments show that in this specific and complex business case AutoML does not achieve as good performances as Ensemble and only in few cases it achieves similar performances of ANN, nevertheless it reduces a lot the time and costs of project.

Considering the analysis executed, we cannot state that AutoML does not achieve good performances because we tested few tools, compared to the number of systems in the market today, with a very specific problem involving many data from many different sources. Since we have a narrow perception of these tools' behavior in different contexts, further experiments are necessary to understand the accuracy and reliability of these tools. The researches of the next future will regard different kinds of problems, probably the next experiments will be on a classification problem. These future researches will be useful to expand knowledge about the performances of AutoML. Even these future experiments will be focused on the comparison between the results obtained by ML solutions and the performances obtained by the AutoML systems. With these new experiments will be used other types of KPIs and metrics to evaluate the goodness of the models developed by the different tools, and certainly will be tested others AutoML systems open source and proprietary to enlarge the benchmark developed during this thesis.

In conclusion, AutoML turned out as a product/service with a very high potential able to shape the companies' business and to improve their data-driven approach. AutoML is still in the early stage of its lifecycle but from the market research was highlighted that it is a continuous growing market, in which innovation and improvement are growing exponentially.

BIBLIOGRAPHY

Adithya Balaji, Alexander Allen, *Benchmarking Automatic Machine Learning Frameworks*, 17 Aug 2018

Andrew Ng, *Machine Learning and AI via Brain simulations*, Stanford University

C. Vercellis, *Business intelligence: data mining and optimization for decision making*, Editor: Wiley, 2009

E. Alpaydin*, Introduction to Machine Learning*, Editor: MIT press, 2014

Frank Hutter, Lars Kotthoff, Joaquin Vanschoren, *AutoML Book*, *Automated Machine Learning: methods, systems, challenges,* 2019

Geron, *Hands-On ML with Scikit-Learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*, Editor: O'Reilly, 2017

Guyon, I., Bennett, K., Cawley, G., Escalante, H.J., Escalera, S., Tin Kam Ho, Macia, N., Ray, B., Saeed, M., Statnikov, A., Viegas, E.: *Design of the 2015 Cha-Learn AutoML challenge*, July 2015

Isabelle Guyon and Lisheng Sun-Hosoya and Marc Boulle and Hugo Jair Escalante and Sergio Escalera and Zhengying Liu and Damir Jajetic and Bisakha Ray and Mehreen Saeed and Michele Sebag and Alexander Statnikov and Wei-Wei Tu and Evelyne Viegas, *Analysis of the AutoML Challenge series 2015-2018*

Marc-André Zöller, Marco F. Huber, *Survey on Automated Machine Learning*, 26 Apr 2019

POLIMI (B. Pernici, C. Francalanci, A. Gerozzano, L. Polidori), IDC (G. Cattaneo and H. Schwenk), JSI (M. Grobelnik), ATOS (T. Pariente, I. Martinez), GUF (T. Ivanov), SINTEF (A. Berre), *DataBench industry requirements with benchmark metrics and KPIs version 1.0*, 2018

QianwenWang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J. Smith, Kalyan Veeramachaneni, Huamin Qu, *ATMSeer: Increasing Transparency and Controllability in Automated ML*, 13 Feb 2019

Quanming Yao, Mengshuo Wang, Yuqiang Chen, Wenyuan Dai, Yi-Qi Hu, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, Yang Yu, *Taking the Human out of Learning Applications: A Survey on Automated Machine Learning*, 17 Jan 2019

Slides of Marketing and Innovation course, Politecnico di Milano, Blue ocean and red ocean strategy (2019)

Slides of the course of Digital Technologies, Politecnico di Milano, Big Data and Data analysis (2019)

Slides of Kimberly Hermans and Jan Mulkens, *Enabling Citizen Data Scientist with Microsoft* (PDF)

T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Editor: Springer, 2011

Thiloshon Nagarajah and Guhanathan Poravi, *An Extensive Checklist for Building AutoML Systems*, May 2019

Thiloshon Nagarajah, Guhanathan Poravi, *An Extensive Checklist for Building AutoML Systems,* 14 April 2019

W. Chan Kim, Renée Mauborgne, *Blue Ocean Startegy*, Harvard Business Review Press, 2004

Xin He, Kaiyong Zhao, Xiaowen Chu, Department of Computer Science, Hong Kong Baptist University*, Survey of the State-of-the-Art*, 14 Aug 2019

WEBOGRAPHY

For each AutoML system considered in this thesis is reported in Table 10 its official website link, if they have it, otherwise, in case of open-source tools there is the documentation found on Github.com.

All the link used to make the AutoML customer research are not reported here but are directly listed in Table 11.

Below are listed all the articles and websites consulted about machine learning and automated machine learning (in squared bracket the ones directly cited inside the thesis):

[1] https://en.wikipedia.org/wiki/Big_data

[2] https://simplicable.com/new/data-veracity

[3] https://www.socialmediatoday.com/news/how-much-data-is-generated-every-minute-infographic-1/525692/

[4] https://en.wikipedia.org/wiki/Machine_learning

[5] https://en.wikipedia.org/wiki/Reinforcement_learning

[6] https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai/#50a826bb7be7

[7] https://elitedatascience.com/machine-learning-algorithms

[8] https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

[9] https://en.wikipedia.org/wiki/Feature_selection

[10] https://medium.com/@lotass/machine-learning-what-you-need-to-know-about-model-selection-and-evaluation-8b641fd37fd5

[11] https://en.wikipedia.org/wiki/Hyperparameter_optimization#Gradient-based_optimization

[12] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#241427cd6f63

[13] https://ai.googleblog.com/2018/01/the-google-brain-team-looking-back-on.html

[14] https://epistasislab.github.io/tpot/

[15] https://www.techopedia.com/definition/28177/data-scientist

[16] https://www.kdnuggets.com/2019/03/why-automl-wont-replace-data-scientists.html

[17] https://www.kdnuggets.com/2019/09/automated-machine-learning-just-how-much.html

[18] https://analyticsindiamag.com/does-automl-work-for-all-data-science-stakeholders-expectations-vs-reality/

[19] https://www.informationweek.com/strategic-cio/team-building-and-staffing/is-automl-the-answer-to-the-data-science-skills-shortage/a/d-id/1334637

[20] https://www.datanami.com/2019/08/28/automl-tools-emerge-as-data-science-difference-makers/

[21] https://hbr.org/2019/10/the-risks-of-automl-and-how-to-avoid-them

[22] https://kortical.com/articles/data-scientists-rejecting-automl

[23] https://www.forbes.com/sites/forbestechcouncil/2019/10/11/implementing-automated-machine-learning-automl/#1b2ceabe554f

[24] https://www.quora.com/Does-Google-AutoML-eliminate-the-need-for-ML-specialists-I-mean-are-they-not-needed-anymore-Im-asking-this-question-because-I-wanted-to-make-my-career-in-ML-deep-learning

[25] https://blogs.gartner.com/carlie-idoine/2018/05/13/citizen-data-scientists-and-why-they-matter/

[26] https://www.forbes.com/sites/forbestechcouncil/2019/02/20/empowering-the-citizen-data-scientist/#113f5a864861

[27] https://go.forrester.com/blogs/who-who-who-are-you-citizen-data-scientist/

[28] http://www.businessdictionary.com/definition/innovation.html

[29] http://www.businessdictionary.com/definition/strategy.html

[30] https://en.wikipedia.org/wiki/Strategy

[31] https://cloud.google.com/automl/?hl=it

Other websites consulted during the researches:

1. http://www.intelligenzaartificiale.it/machine-learning/

2. https://en.wikipedia.org/wiki/Automated_machine_learning

3. https://en.wikipedia.org/wiki/Machine_learning

4. http://www.intelligenzaartificiale.it/machine-learning/

5. https://ai.googleblog.com/2018/01/the-google-brain-team-looking-back-on.html

6. https://elitedatascience.com/machine-learning-algorithms

7. https://en.wikipedia.org/wiki/Machine_learning

8. https://hackernoon.com/a-brief-overview-of-automatic-machine-learning-solutions-automl-2826c7807a2a

9. https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f

10. https://medium.com/@jrodthoughts/understanding-semi-supervised-learning-a6437c070c87

11. https://medium.com/@lotass/machine-learning-what-you-need-to-know-about-model-selection-and-evaluation-8b641fd37fd5

12. https://medium.com/@ODSC/the-past-present-and-future-of-automated-machine-learning-5e081ca4b71a

13. https://ml.informatik.uni-freiburg.de/papers/15-NIPS-auto-sklearn-preprint.pdf

14. https://towardsdatascience.com/everything-you-need-to-know-about-automl-and-neural-architecture-search-8db1863682bf

15. https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

16. https://www.datarobot.com/platform/what-is-automated-machine-learning/

17. https://www.datarobot.com/wiki/automated-machine-learning/

18. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#241427cd6f63

19. https://analyticsindiamag.com/does-automl-work-for-all-data-science-stakeholders-expectations-vs-reality/

20. https://www.informationweek.com/strategic-cio/team-building-and-staffing/is-automl-the-answer-to-the-data-science-skills-shortage/a/d-id/1334637

21. https://www.quora.com/Does-Google-AutoML-eliminate-the-need-for-ML-specialists-I-mean-are-they-not-needed-anymore-Im-asking-this-question-because-I-wanted-to-make-my-career-in-ML-deep-learning

22. https://www.kdnuggets.com/2019/03/why-automl-wont-replace-data-scientists.html

23. https://www.forbes.com/sites/forbestechcouncil/2019/10/11/implementing-automated-machine-learning-automl/#1b2ceabe554f

24. https://towardsdatascience.com/automl-for-data-enthusiasts-30582b660cda

25. https://www.datarobot.com/wiki/citizen-data-scientist/

26. https://www.forbes.com/sites/forbestechcouncil/2019/05/24/machine-learning-for-the-rest-of-us-how-to-cultivate-the-citizen-data-scientist/#3baedc054b0f

27. https://www.forbes.com/sites/stevebanker/2018/01/19/the-citizen-data-scientist/#16fdf2512702

28. https://www.silicon.it/data-storage/business-intelligence/gartner-descrive-il-citizen-data-scientist-102500

AKNOWLEDGMENT

Con la stesura di questa tesi si conclude il mio percorso universitario e per concluderlo al meglio mi sento in dovere di ringraziare tutte le persone che ho sentito vicine in questi anni e senza le quali tutto ciò che ho fatto sarebbe stato letteralmente impossibile.

Vorrei innanzitutto ringraziare la professoressa Barbara Pernici per avermi seguito e consigliato durante la stesura della tesi e per avermi fatto scoprire e amare il mondo digital con il suo corso di Digital Technology, il quale mi ha orientato a sviluppare una tesi rivolta ad una delle più interessanti innovazioni nel campo machine learning e data science: i sistemi AutoML. Inoltre, devo un generale ringraziamento a tutti i professori che durante questi anni con i loro insegnamenti mi hanno trasmesso metodo e conoscenze professionali.

Allo stesso modo devo ringraziare Bip (Business Integration Partners) per avermi permesso di sviluppare la tesi internamente all'azienda, sono stati sei mesi in cui ho imparato tanto e ho avuto l'occasione di conoscere tanti esperti nel mondo della data science e della consulenza. Andando più nello specifico ringrazio Bip.xTech, il centro di eccellenza di data science di Bip, dove ho trascorso il mio periodo di stage in cui ho avuto il piacere di trovare un gruppo di lavoro davvero fantastico, il Team Fox, che mi ha fatto sempre sentire parte del team e mi ha aiutato e consigliato nello sviluppo della tesi. In particolare, vorrei ringraziare il mio correlatore Alessandro Volpe, che mi ha aiutato dall'inizio alla fine con la sua esperienza nei momenti di difficoltà incontrati durante lo svolgimento. Un grazie anche a Gerry, Enrico, Matteo, Davide, Carlo e Giovanni per avermi aiutato quando avevo un dubbio da chiarire o semplicemente per una chiacchiera davanti a un caffè.

Grazie a tutti per la mia prima esperienza in azienda, davvero stupenda.

I cinque anni di università sembrano volati adesso che scrivo i ringraziamenti della tesi, ma pensandoci bene non lo sono stati affatto. Sono stati anni con tanti alti e bassi. Ma tutte le esperienze fatte sarebbero state prive di significato se al mio fianco non avessi avuto i miei compagni di avventura. Edo e Aldo, ci siamo seduti a fianco il primo giorno di lezione per puro destino, il resto lo sapete anche voi. Fede mi hai sempre trasmesso voglia di fare e di non arrendersi davanti a niente. Albi sei uno tenace, altruista e su cui posso sempre contare. Sara sei un uragano di energia, spero di avertene rubata una minima parte che mi servirà da qui in avanti. Ringrazio tutti di cuore. Siete dei veri amici.

Devo ringraziare di cuore due persone speciali che mi sono sempre state vicine e con le quali ho condiviso non so quante esperienze da cinque anni a questa parte. Mattia grazie per esserci sempre stato sia in momenti difficili che in momenti di puro divertimento, so che su di te potrò sempre contare, sei un caro amico. Erica sei la persona con la volontà più ferrea che ho mai conosciuto, un esempio da prendere quando si vacilla di fronte alle difficoltà. Siete sempre stati presenti per me, vi voglio bene.

Elena con te ho condiviso tutte le mie paure, gioie ed emozioni di questi anni. Sei la prima persona a cui mi rivolgo per ogni minima cosa e che mi capisce subito, basta solo uno sguardo tra di noi per dirci tutto. Mi hai incoraggiato nei momenti difficili e mi sei stata sempre vicina dandomi tutto il tuo amore, non posso che ringraziarti dal profondo del cuore.

Un grazie sincero alla mia sorellina Teresa per avermi sempre spronato, anche indirettamente, a fare di più e a non accontentarmi. Siamo cosi diversi ma anche cosi uguali, io calmo te impulsiva, io pigro te energica, *io intelligente te mica tanto,* io pacifico te guerriera e potrei andare avanti ancora tanto. Continuiamo a completarci l'un l'altro che sarà la nostra carta vincente per il futuro.

A mio Papà devo l'uomo che sono e che diventerò, il primo a cui dico sia i miei successi che fallimenti. Ti cerco per consigli, per una linea guida da seguire quando sono perso nell' incertezza, per capire il tuo punto di vista più esperto. Sei la mia forza da sempre. Ti voglio bene.

A mia Mamma le dovrei scrivere almeno qualche pagina di ringraziamenti per tutte le cose che fa e che ha fatto per me, qui ti dirò semplicemente grazie per credere sempre in me e aiutarmi con tutte le tue forze in ogni passo che faccio. Sarai sempre il mio porto sicuro. Ringrazio di tutto cuore anche te Pero, ti ho sempre sentito vicino durante questi anni e so che ti sei preso cura di me come se fossi un figlio. Vi voglio bene.

In ultimo vorrei ringraziare due persone fantastiche, due donne favolose ma soprattutto due nonne amorevoli. Un immenso grazie alla mia Nonna Tina per avermi sempre amato con dolcezza e per essermi sempre stata vicina. Infine, ringrazio mia nonna Rina, la persona che più di tutte ha creduto in me e mi ha fatto capire quanto importante sia lo studio e l'impegno per raggiungere i propri traguardi.

Grazie.