Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Diploma Thesis

Predictive modelling of customer value

Michal Zawal

© 2020 CULS Prague

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Michał Zawal

Systems Engineering and Informatics
Informatics

Thesis title

Predictive modelling of customer value

Objectives of thesis

This thesis helps to determine the factors affecting customer behaviour. It analysis the connection between several independent attributes, such as personal and macro economical to the success of sales. The aim of the research is to create a prediction model which, based on historical transactions, will be able to select the key attributes for effective sales. In this way, by analyzing the attributes of a new client, the model will be able to predict the probability of sale, estimating its prospective value. In such a way, it will contribute to the classification of customers based on their value and ultimately to increasing the profitability of sales.
The data is a result of a direct marketing campaign performed by a Portuguese banking institution to sell term deposits/certificate of deposits. The banking institution made phone calls to potential buyers from May 2008 to November 2010.

Methodology

The methodology used in the study assumes statistical analysis and comparison of the influence of macroeconomic and behavioural factors mentioned in the literature review with those obtained in the modelling process.

During the research the term big data analysis was introduced, presenting the problem of a large amount of data and showing what problems can be solved and how it brings profit for the company.

During the practical part of the study, the data will be presented in descriptive statistics, then standardised and cleaned to be usable for modelling. During the modelling process, models using classical logistic regressions as well as machine learning algorithms will be created to select the model with the highest prediction accuracy.

The proposed extent of the thesis

60 – 80 pages

Keywords

Statistical analysis, predictive modelling, customer analysis, machine learning, big data

---

Recommended information sources

ABBOTT, D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.

ATHANASOGLOU, P., BROSSIMIS, S, DELIS, M. "Bank-specific, industry-specific and macroeconomic determinants of bank profitability", Int. Fin. Markets, Inst. and Money 18 (2008).

FORTE, R. Mastering Predictive Analytics with R. Praha: Packt Publishing, Limited, 2015. ISBN 9781783982813.

GODDARD, J., LIU, H. MOLYNEUX, P., WILSON, J. "The Persistence of bank profit", Journal of Banking & Finance vol. 35 (2011).

JOBBER, D., LANCASTER, G : Selling and Sales Management. Harlow: Pearson Education Limited, 2009. ISBN: 978-0-273-72065-2.

KOTLER, P. – ARMSTRONG, G. *Principles of marketing.* Harlow: Pearson, 2012. ISBN 978-0-273-75243-1.

LINOFF, G. – BERRY, M J A. *Data mining techniques : for marketing, sales, and customer relationship management.* Indianapolis: Wiley, 2011. ISBN 978-0-470-65093-6.

PASIOURAS, F., KOSMIDOU, K. "Factors influencing the profitability of domestic & foreign commercial banks in European Union", Research in International Business and Finance, vol.21, no.2, (2007) pp. 222-237.

TURHANI, A., HODA, H., "The Determinative Factors of Deposits Behavior in Banking System in Albania", Academic Journal of Interdisciplinary Studies Vol 5, No 2, (2016).

ZAWADI, A. "Determinants of Banks' Profitability in a Developing Economy: Empirical Evidence from Tanzania ", European Journal of Business and Management Vol.6, No.31 (2014).

Expected date of thesis defence

2019/20 SS – FEM

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 11. 11. 2019

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 12. 11. 2019

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 09. 02. 2020

Declaration

I declare that I have worked on my diploma thesis titled "Predictive analytics of customer value" by myself and I have used only the sources mentioned at the end of the thesis. As the Predictive modelling of customer value author of the diploma thesis, I declare that the thesis does not break copyrights of any their person.

In Prague on _____

Acknowledgement

I would like to sincerely thank Ing. Tomas Hlavsa, Ph.D. for guiding me through the rough path of data analysis journey, supporting and advice my ideas with his extensive knowledge and experience and very cooperative attitude, my parents for giving me the opportunity to be in that place where I'm now and all of my relatives, who were keeping their fingers crossed for me.

Predictive analytics of customer value

Abstract

The presented thesis aims to present a way of data analysis, and more precisely, predictive analytics can be applied in many fields of science, knowledge development and the process of gaining a competitive advantage. The main objective of the research is to apply statistical models of supervised machine learning in an attempt to identify the consumer characteristics that have the greatest influence on purchasing decisions. These studies are expected to have a positive effect both on the company using these techniques through appropriate customer targeting, but also on the customers themselves, as the product will be able to reach the interested audience.

The impact of the indicators of the prediction model will also be compared to formerly conducted studies of other analysts. Besides merely personal characteristics, the impact of the macroeconomic environment in which decisions were made and the history of consumer behavior will also be assessed. Additionally, before building the model, the customer data set will be well examined in detail using univariate and multivariate methods of statistical analysis. The data will be then cleaned and prepared for the model construction.

Keywords: Machine Learning, Predictive Analytics, Big Data Analytics, Statistical analysis, Data Science, Artificial Intelligence, Decision Tree, Logistic Regression, CRISP-DM, Supervised Learning, Customer targeting.

Predictive Analytics hodnoty zákazníka

Abstrakt

Předkládaná práce si klade za cíl představit způsob analýzy dat a přesněji prediktivní analýzu lze aplikovat v mnoha oblastech vědy, rozvoj znalostí a procesu získání konkurenční výhody. Hlavním cílem výzkumu je aplikovat statistické modely učení pod dohledem ve snaze identifikovat charakteristiky spotřebitele, které mají největší vliv na rozhodování o nákupu. Očekává se, že tyto studie budou mít pozitivní dopad na společnost využívající tyto techniky prostřednictvím vhodného cílení na zákazníky, ale také na samotné zákazníky, protože produkt bude schopen oslovit zájemce. Dopad indikátorů predikčního modelu bude rovněž porovnán s dříve provedenými studiemi jiných analytiků. Kromě čistě osobních charakteristik se bude posuzovat i dopad makroekonomického prostředí, ve kterém byla učiněna rozhodnutí, a historie chování spotřebitelů. Před sestavením modelu bude navíc podrobně prozkoumána sada zákaznických dat pomocí statistických analýz univariate a multivariate. Data budou poté vyčištěna a připravena pro konstrukci modelu.

Klíčová slova: Strojové učení, Predictive Analytics, Velká data, Statistická analýza, Data Science, Umělá inteligence, Rozhodovací strom, Logistická regrese, CRISP-DM, Dozorované učení, Cílení na zákazníka.

Contents

1.Introduction

Nowadays, due to the big data revolution the amount of data produced and stored is growing exponentially. The change includes as well possible sources of gathering the data, currently it's available from sources, completely unexpected years ago. This allowed us to start recording, storing and then analyzing data from such sources as internet activity, including social networking entries and search histories, GPS signal of mobile devices or satellite image processing. The Big Data market is still relatively young, but is growing very dynamically. As the number of companies providing this service increases, so does the perception of companies which notice that the use of advanced data analysis is not only an additional bar for the management, but mainly something that is able to help them gaining an operational competitive advantage. Data Science services allow for better targeting of customers, increasing efficiency of supply chains or identification of payment frauds.

So far, companies from many industries have so far decided to implement Big Data applications. As examples of successful solutions we can mention Netflix. The American streaming platform providing entertainment content carefully watches what its customers are viewing, and then, using predictive analytics, decides to deliver and produce series that best match the current trends. Netflix also developed a recommendation system to perfection, targeting content to different audiences[1]. Another example of company development based on big data technology is General Electric. In its "Industrial Internet" project, the energy producer used IoT technologies, including numerous sensors monitoring the work of the machine infrastructure to control temperature, fuel level and failures. In this way, interruptions in energy production were reduced, thus reducing the loss of revenue[2]. Another example of Big Data application is the American company SpaceKnow[3]. SpaceKnow provides services in the form of satellite map analysis, object identification and their dynamics over time. This allows to perform geospatial analysis. As we can observe the Big Data market provides many possibilities and offers many analyses. Their applications and uses also expand over time. No surprisingly,

[1] https://neilpatel.com/blog/how-netflix-uses-analytics/
[2] https://www.ge.com/digital/blog/everything-you-need-know-about-industrial-internet-things
[3] https://spaceknow.com/

the global big data & business analytics market is expected to grow from $168.8 billion in 2018 to 274.3 billion by 2022[4].

Data revolution brought opportunities as well as threats. It's is affecting human's activity, along with concepts of smart cities, intelligent agents and autonomous vehicles there are player that are trying to build data science solutions purely for commercial profit. One should note that this undoubtedly growing industry will affect everyday life more and more significantly, thus it should be utilizing this possibilities ethically and sustainable, giving a chance to address obstacles which current world is facing.

2. Objectives and methodology

2.1 Objectives of the thesis

The main objective of the work is to perform predictive analysis in the chosen set of data. Predictive analysis will allow to address the needs of a bank conducting marketing campaigns on sales of deposits. Predictive analytics can support bank in its campaign by optimizing its operational activity, reduce costs and finally acquire more customers. The aim of the bank's marketing department is to identify customers who are more willing to purchase deposits. Therefore, during the theoretical part, we will try to review concepts of big data analysis as well as results of research conducted by many analysts on the impact contributed by particular attributes. Then, in the practical part, the goal is to build a model which, on the basis of the input data, will select those customers who are most willing to buy bank deposits.

In its theoretical part, the paper introduces concepts and utilizations of Big Data, Data Mining, Predictive Analytics with the emphasis on impact it could have on modern enterprises. Theoretical part of the paper also presents studies of consumer behavior and factors, including those not directly related to the client, but might influence buyer's purchasing decisions. This section also presents the foundations, the processes and methods used for statistical data

[4] https://www.cognetik.com/blog/data-analytics-market-in-2020-trends-forecasts-challenges/

analysis, as well as the descriptions of the most popular machine learning algorithms used nowadays.

The practical part focuses on the exploration the data and construction of a predictive models including all necessary steps required for this purpose. The goal of modelling is to enable discovery and explanation of meaningful patterns impacting consumer decisions. The aim of the research on consumer behavioral relationships is to optimize the effectiveness of the business through the classification and following target marketing, that is ultimately increasing the profitability of sales. The data set used in practical modelling has been created as a result of a direct marketing campaign conducted by a Portuguese. The data set consist of information of sale performance coded as dummy variable (success/failure), and the input variables including personal, macro economical and customer's history . In order to achieve the highest accuracy, different predictive modelling methods are being used to provide the most satisfying model.

2.2 Methodology

The methodology of the study assumes carrying out a predictive analysis on a set of data gathered during a marketing campaign of a Portuguese bank aimed at selling its deposits. The data set includes 41 188 observations containing an explanatory variable and 20 independent explanatory variables. Methodology of presented paper involves conducting researches with the usage of CRISP-DM method, assuming the following steps to be undertaken:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation

The first stage of the research will aim to understand how business can benefit from the use of predictive analytics and identify what type of problem can be adressed and hopefully resolved by model deployment. The second stage assumes detailed examination of the data set in respect of quality of possess data and contained outliers. Data understanding phase will be underdone with a usage of univariate and multivariate descriptive statistics. The goal of this phase is to

learn more about the distribution of numeric variables as well as frequencies of categorical ones. The following phase, data preparation is aimed to correct flaws, previously discovered and prepare the data set for the modelling. This stage also allows to assess the quality of the data and the errors that should be corrected. Next phase, modelling assumes development of predictive model on the top of previously polished data set. Finally the evaluation phase is the moment when previously build model will be assess, whether it predictive capabilities reached goals assumed by research. The evaluation moment is a point when developed model will receive feedback, whether it's worth to deploy it or it should be rebuild.

3. Literature Review

3.1 Introduction to concept of Big Data Analysis

"An analysis, I suppose, may be thought of as a kind of breaking down or decomposing of something. So we have the picture of a kind of intellectual taking to pieces of ideas or concepts; the discovering of what elements a concept or idea is composed and how they are related. Is this the right picture or the wrong one—or is it partly right and partly wrong? That is a question which calls for a considered response." (Strawson, 1992)

Before examining more complex terms, it's valid to bring the term "analysis" closer. Analysis is the process of examining data and facts to uncover patterns and understand the cause-effect relationships that form the basis for problem solving and decision making. Term of analysis has been applied during the study of mathematics and logic by Aristotle (384-322 B.C.) and comes from Ancient Greek word *ἀνάλυσις* (*análisis*, "a breaking-up" or "an untying;" from *ana-* "up, throughout" and *lisis* "a loosening" (Online Etymology Dictionary, 2019)

Although the term of analysis is not new, the meaning, that is currently associated with rose in 2005 due to introduction of Google Analytics. Nevertheless operations related with analysis has been performed long time before the technological giant from California was founded. The analysis, in current terms has been the foundation of modern science in recent centuries. The world of science has been formulating and solving new problems by doubting, hypothesizing and experimenting on the data to finally gather knowledge. Besides of its origin in mathematical studies analysis has been also being widely used in areas such as chemistry, biology, physics, economics, engineering, music, psychology and statistics. Today, it is important, as it has never

been before. Analysis plays an extremely important role in the world of modern business, allowing to prove suspicions and understand the cause - effect sequence, allowing the mining of information to support decision-making. Current business world has found many ways of gaining from analytics. Various disciplines, departments and specializations has arisen in bypast years the enabling business development. Departments such as: data analysis, software analysis or process analysis can be distinguished in different disciplines.

This paper will focus on data analytics with the usage statistical methods to analyze big data sets gathered by business environment, during so called big data analysis. In big data analysis choice of framework depends on the objective for which the information is intended to be gathered. The objectives of data analysis may vary, as may the application of the acquired knowledge. Nowadays, data analysis can be used for purposes such as:

1. Predicting (whether a credit card transaction is fraudulent or not),
2. Finding patterns inside the data (tracking clicks on websites).
3. Discovering data relations (Finding and grouping similar media articles).

On the basis of the differences in objectives and methodology, we are able to distinguish 4 types of conducted analyses:

1. Descriptive analytics - Provides a way to analyze and summarize historical data to provide an easier way to interpret them. Descriptive analytics tries to answer the question - What happened? Methods of descriptive analysis include summary statistics describing central tendency, spread, shape of distribution and shape. Descriptive statistics help to provide a greater insight into the data and to present them in a summarized form. Examples of measures of descriptive statistics are min, max, mean, median, std deviation, kurtosis.
2. Diagnostic analytics - Assumes analysis of historical data to discover the reasons why particular events occurred. Diagnostic analytics tries to answer questions - *why did it happen*? While descriptive analytics, calculating different statistics tries to measure what happened, diagnostics analytics using data mining tools tries to find relationships and patterns inside the examined data. Examples of tools used in data mining are statistical learning methods such as clustering and dimension reduction.

3. Predictive analytics - comprises prediction of the occurrence of a given event or estimation of the probability of its occurrence or prediction of future values through the usage of models. Predictive analytics tries to answer the question - *What is likely to happen?* One example is to predict whether tumor is benign or malignant. Predictive models are being built by the usage of already existing historical data. These models learn and adjust their parameters to match the patterns and trends from existing data. Predictive modelling include such a methods like classifications and regressions. For this purpose, such tools as Logistic Regression, Neutral Networks or Decision Trees are used.
4. Prescriptive Analytics – While predictive analytics attempts to predict the outcome of an event, prescriptive analytics goes one step further and using multiple predictive models simultaneously in order to predict different results and their interactions. Prescriptive analytics tries to answer the question - *What can we do to make it happen?* Prescriptive analytics being a kind of what-if analysis tries to predict possible results based on different states of choices. The most commonly used method in prescriptive analytics is optimization. (Bahga & Madisetti, 2019)

Today's technology, framework, computing power and rapidly growing availability of data present an entirely new opportunity and challenges for analysis. Current data collection systems such as sensors and cameras allow to obtain data without human interaction, thus increasing the availability of data to a level that has never been seen before and is still expected to growth. Today, decision making systems are able to use this data in a fraction of a second. Data analysis has an impact on ubiquitous technologies in today's world, significantly affecting the functioning of societies, and their role will certainly grow in the future, as will the number of data sources and algorithms used to analyze them.

3.1.1 Big Data

The data in traditional meaning is the quantiles, characters or symbols on which operations are performed by a computer which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical or mechanical recording media. As information technology has evolved, including the Web 2.0, revolutions of smartphones and the Internet of Things, the

number of sources of data has risen significantly and with this form of obtained data has significantly changed its structure, allowing you to receive and process data in forms such as GPS location, camera view or even sounds data. Along with the emergence of new data sources, a description of the following term was also created.

According to Frank J. Ohlhorst *Big Data* defines a situation in which data sets have grown to such enormous size that conventional information technologies can no longer effectively handle either the size of data set or the scale and growth of the data set and further gathering value out of it and managing it is hardly impossible. Ohlhorst provide that main difficulties with Big Data appears in the sections of acquisition, storage, searching, sharing, analytics and visualization data. (Ohlhorst, 2013). Additional definition of *Big Data* is coming from Oracle states that Big Data is larger, more complex data sets, especially from new data sources, which are so voluminous that traditional data processing software just can't manage them. But additionally with an ability to address business problems have never been wouldn't able to tackle before. (Oracle, 2019)

Within the types of Big Data we can distinguish three forms:

- Structured - Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.
- Unstructured - Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.
- Semi structured – Consisting of both, structured and unstructured data simultaneously.

Bahga & Madisetti has distinguished five characteristics used to described Big Data, thus called 5V's as follow:

- Volume - The volume of data refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities.
- Velocity - The term refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.
- Variety – Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. To obtain information from data in this format, Big Data systems must be able to process data from these formats.
- Veracity – Veracity refers to the quality of the data that is being analyzed. High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise.
- Value – This term refers to the applications and intrinsic value of the data. The goal of each Big Data system is to gain some knowledge from the data. The data value can also be described by its accuracy. For some systems, the velocity of data processing will also be crucial. (Bahga & Madisetti, 2019)

Having the term and characteristics of Big Data enclosed, it's really critical to indicate why analysis of Big Data really matter. Today's Big Data systems are able to obtain, transform and use data that has never been used before. This allows us to obtain information that is unavailable and even unnoticeable until now. Based on system requirements it can provide with batch or

real-time information delivery for various types of analysis including descriptive, diagnostics, predictive or prescriptive. The examples of usage the Big Data analytics are listed as follow:

- Stock exchange data,
- Jet Engine's performance data analyzed in real time during flight time,
- Data coming from social media consisting of text, images, audio and video data,
- Sensors included in autonomous vehicles systems,
- Transactions from banking and financial applications,
- User's click stream generated by web applications,
- Smart Parking, making finding parking spots easier for users,
- Early forest fire detections,
- Water Quality monitoring.

Authors Bagda & Madisetti pointed out that Big Data has the potential to backup next generation of smart applications making them and more intelligent leading to develop various domain of business applications and current society life's including surveillance visions, environment, Internet of things, retail and many more. (Bahga & Madisetti, 2019).

3.1.2 Data Mining

Data mining is the process of transferring data into information, gathering knowledge and finding useful patterns within the data,. This process starts with data, which can be a set of several observations or a matrix consisting of a million rows and a couple of thousands of columns. The data mining process uses specialized methods to extract useful knowledge. These methods come from disciplines such as statistics and machine learning. (Kotu & Deshpande, 2015)

Although the popularity of Data Mining has increased in recent years, it is not a brand new term, and most techniques have existed for decades, at least in academic environments. Nevertheless, the noticeable increase in the use of Data Mining is currently mainly the result of the following factors:

- Significantly more data is produced – The value of the analysis increased with the increase in the amount of available data. To some extent, Data mining requires a large amount of data in order to train and test models.

- The data that is produced are being stored - multiple available data sources were not taken into consideration in the analysis context and were therefore not stored. As the variety of data being analyzed grew, data in the form of audio, images or films began to be stored..
- Increased attention to the opportunities offered by Data Mining- Among many various disciplines, companies realized that performing an accurate Data Mining analysis provides knowledge to help them gain a competitive advantage.
- Software for analysis is more user-friendly and available – Thanks to the development of data storage technology and lower computing costs, Data Mining has become cheaper and more accessible. With the growing interest in the results of analyses, Data Mining methods have been commercialized, thus allowing people from outside the academic circles to use it. (Linoff & Berry, 2011)

Data Mining being a process is not composed of a single operation. There is no universal instruction, but in order to increase the safety of the analysis, stop the data mining being performed haphazardly and avoid reinventing the wheel, a certain standard was required.

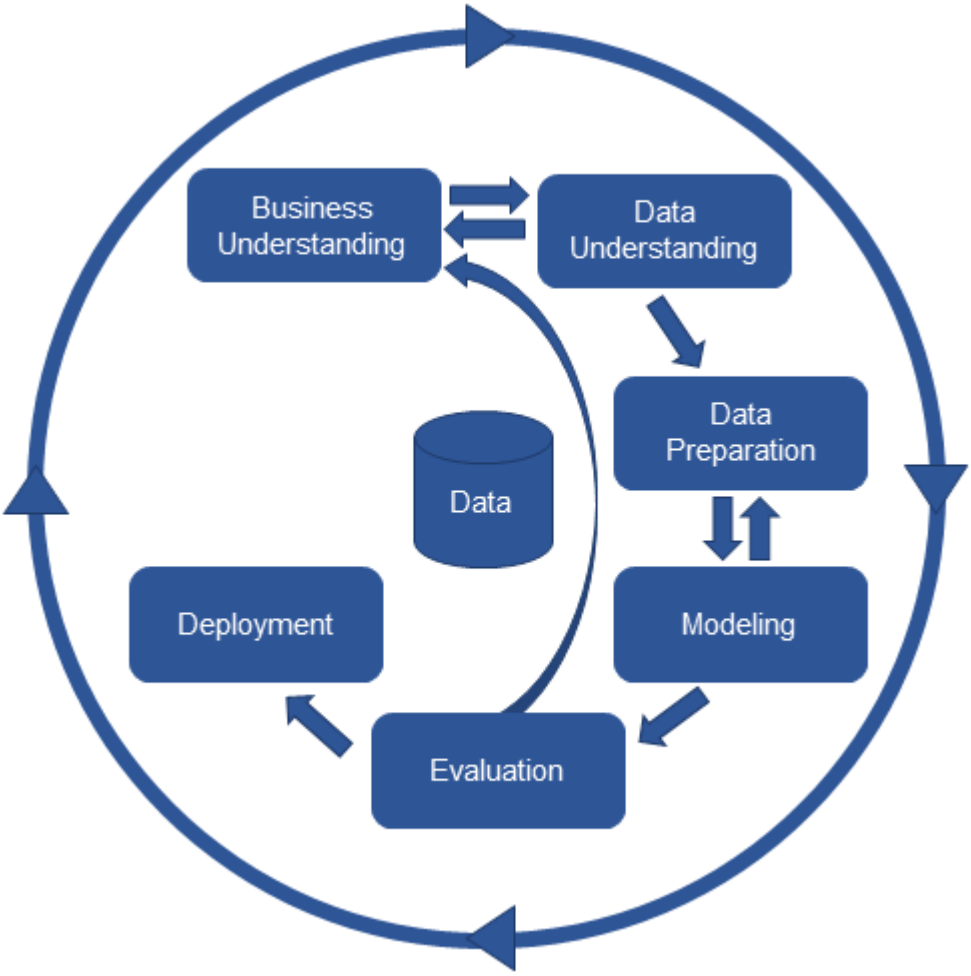

Figure 1 CRISP-DM process

Source: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

The Cross-Industry Standard Process for Data Mining (CRISP-DM) has been created as industry-neutral, tool-neutral and application neutral. According to CRISP-DM, the single Data Mining project consists of 6 phases, visible in figure 1. It is important to note that the phase sequences in this process are dependent on each other, i.e. movement between phases is possible in the directions indicated by the arrows. In addition to the obvious forward movement, the arrows also allow the return to the previous phases when the characteristics and behavior of the model require it. After the actual modelling phases, the results are being evaluated. If the results are not satisfactory at this point in time, the model returns to the previous phases. (Larose & Larose, 2015)

Process CRISP-DM can be divided into 6 phases:

1. Business Understanding Phase – In this phase, the project objectives and business requirements are formulated. Then the requirements are translated into a problem in the context of data mining. Finally, reaching to formulate preliminary strategy.
2. Data Understanding Phase - First, it implies collecting the data, then performing an exploratory data analysis in order to get familiar with the data and to have the first insight into it. The preliminary insight into the data also makes it possible to evaluate the quality of the data.
3. Data Preparation Phase – This phase involves all aspects of the preparation of the data from the raw, basic to the final form, which will be used in the following phases. In this phase, the variables and observations are prepared to fit the scope of the analysis. If necessary, the transformation of the variables is also done.
4. Modelling Phase – In this phase, the modelling techniques are selected and the actual modelling is carried out. Often several techniques are used in the same case to find the most effective solution. At this stage it is possible to return to the previous phase of Data Preparation in order to calibrate the data more accurately to match the particular modelling technique..
5. Evaluation Phase – After the modeling phase has brought a model, it needs to be evaluated in terms of its quality and effectiveness before it can be implemented. At this stage, it is compared whether the model has achieved its objectives in the first phase. Ultimately, this phase determines whether the built model will be used in practice.
6. Deployment Phase – The successful assessment of the model is followed by the implementation phase. For this purpose, it shows how the model is able to solve research/business problems.

Knowing how the Data mining project is running, it is worth to specify what kind of tasks can be performed using it. Authors D. Larose and C. Larose listed 5 popular tasks carried out using Data Mining:

- Description – This process is about finding a way to explain what is inside the data. The description of trends and patterns often suggests their possible origin. In order to

describe data the Data Mining model should be as transparent as possible. Descriptive analytics can be done through the usage of summary statistics and explonatory data analysis.

- Estimation – the numerical value of the target variable is estimated on the basis of a set of collected numerical and/or categorical data. Building a model on historical data is based on estimating its parameters so that they reflect the influence of independent variables on the target variable to the greatest extent possible. Then, when a new variable appears in the model, its target variable is estimated on the basis of the variables describing it. One of simpest method being used for estimation purposes is linear regression.
- Classification – In particular is similar to the estimation, with the difference that the target variable is a categorical variable instead of a numerical variable as in the case of estimation. The classification can assign the target variable to a binary variable or to specific categories.
- Prediction - is similar to estimation and classification, with the exception that the prediction results are located in time. Prediction tasks may include predicting the value of stocks in 3 months.
- Clustering – refers to the grouping of observations into classes of similar objects. Cluster is a collection of similar observations. Clustering differs from classification in that there is no target variable here. The task of clustering is not to classify, estimate or predict target variable values, instead clustering algorithms look for a homogenous sub-group.
- Association – is to find out which variables are mutually dependent. The purpose of the association is to discover how variables influence each other, how strong this influence is, and what is the direction of these relationships. A typical task of an association is to analyze customer basket. The association's rules are in the form of: *if the precedent then the following*, together with measures of support and confidence of the association. (Larose & Larose, 2015)

3.1.3 Predictive Analytics

Nowadays, the terms Predictive Analytics and Data Mining are sometimes used interchangeably, and in the period of recent increases in interest in these types of analyses, both terms are often mistakenly defined. Dean Abbott, mentions in his book that since Predictive Analytics became a popular term, has started to use Predictive Analytics and Data Mining interchangeably (Abbott, 2014, p. 13). On the other hand, the authors of Larose & Larose specify predictions as one of the actions performed by Data Mining, indicating that they differ in the way that the result of the prediction is situated somewhere in the future. (Larose & Larose, 2015, p. 12). Intuitively, in my opinion, it is the most relevant definition. To conclude, this means that Predictive Analytics as well as Data Mining use statistical methods and statistical learning algorithms for activities such as estimation, clustering and classification, however, the result of predictive analytics is situated in the future. This means that there are also differences in the applications of the different types of analysis. Comparing this to Bahga & Madisetti *Big Data Analytics*, Data Mining can be compared to Diagnostic Analysis, answering the question *why a given event took place*, indicating patterns and relationships between data. Predictive analytics helps to answer the question of *what may happen* and allows you to estimate the probability of the event.
(Bahga & Madisetti, 2019)

As a result of the above described characteristics of predictive modelling, there is a shift in the access to data used in predictive analytics. This means that the target variable value is not known when the independent variables are known for both the historical and future data, as shown in figure 2. This means that if there was a link between the historical and the present data among the historical data, this link should be maintained.
(Abbott, 2014, p. 10)

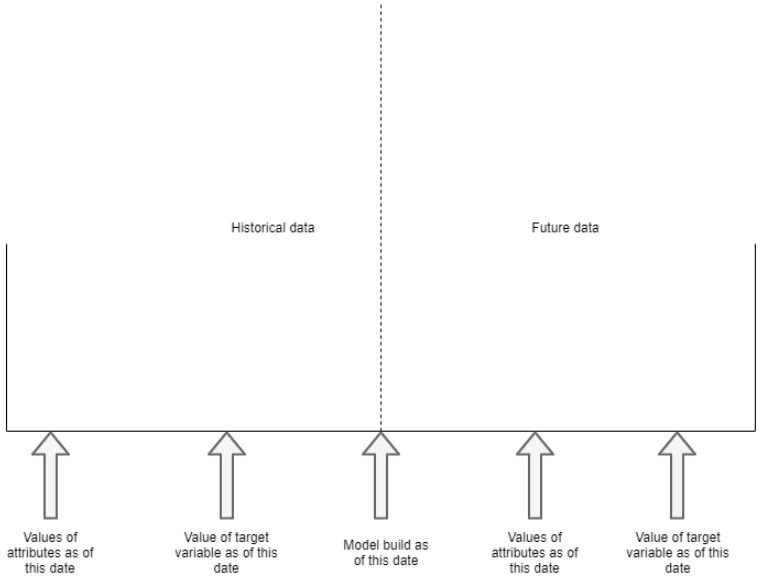

Figure 2 Predictive analytics timeline

Source: Author's own work

In recent years Predictive analytics has successfully found its application in areas such as:

- Analytical customer relationship management
- Clinical decision supporting system
- Cross-sell
- Customer retention
- Direct marketing
- Fraud detection

3.2 Factors affecting customer's decision making

3.2.1 Data set attributes

Having discussed the terms of Big Data, Data Analysis and its different variants and applications, it is possible to focus on the data set that will be analyzed. In this section the theoretical foundations of the impact of independent variables on target variable will be discussed. The analyzed data set was created during the marketing campaign of the Portuguese bank in 2008-2013. The marketing campaign aimed at telephone sales of long-term bank

deposits. The data set collected during the campaign consists of 41 188 observations containing 20 variables explaining the dependent variable. The dependent variable of this data set is a binary variable informing about whether the sale of the deposit was a successful or a failure. The explanatory variables in the data set can be distinguished into 3 types:

- Personal characteristics features
- Macro economical factors
- Client & campaign related features.

In the further part of this subsection, a detailed description of the impact of each attribute on the sale and specifically on the sale of bank deposits will be presented.

3.2.2 Personal characteristics factors

Consumers among societies vary tremendously in age, income, education level, and tastes. They also buy an incredible variety of goods and services. How these diverse consumers relate with each other and with other elements of the world around them impacts their choices among various products, services, and companies. Consumer buyer behavior can be defined, as the buying behavior of final consumers individuals and households that buy goods and services for personal consumption. (Kotler & Armstrong, 2016).

Among the factors influencing consumer behavior there are the several groups that have the greatest influence. P. Kotler and G. Armstrong distinguish 4 main groups of factors influencing customer decisions:

1. Cultural Factors
    a. Culture – Denotes a set of fundamental behaviors, ways of understanding the world, needs, values and behaviors taught by a member of the community from family, environment or other influential institutions.
    b. Subculture – Each culture contains subcultures. Subculture is defined as a group of people sharing value systems on the basis of shared experiences and values. Subcultures can consist of people with a common nationality, religion, ethnic group or geographical region.

c. Social class – These are relatively constant and orderly divides within the society, whose members share similar characteristics, values, behavior and needs.

2. Social Factors
    a. Groups and social networks – Groups consist of two or more individuals and are known to influence the people who are in them. We can distinguish between two types of groups, namely the membership group, which directly affect the members of the people who belong to them, and the reference group, which is the group to which the individuals refers and makes comparisons.
    b. Family – The family is the most important social group in the whole society for the consumer, its welfare and well-being is always taken into account when making decisions. The roles that an individual fulfils in the family and the opinions of other members of the family can be crucial for the customer's decisions.
    c. Roles and Status – Individuals can belong to many social groups as well as online communities. The position held within these groups defines terms such as role and status. The role defines the activities expected from the individual, in the way that people around are expecting from it.
3. Personal Factors
    a. Age and Life Cycle Stage – People change their buying habits during their lives. The type of food, home furnishings and the way they spend their free time is strongly influenced by age. For this reason, purchasing decisions are strongly linked to the moment of life of the consumer and his or her family. The life cycle usually changes with demographic changes and events that have a key impact, such as marriage, the birth of children, the purchase of a home or the moment of retirement.
    b. Occupation – A person's profession has a significant influence on his or her purchasing choices. White collars buy statistically more suits, while Blue-collar workers tend to buy more workwear. In order to increase sales of a product, a company should identify a professional group with an above-average interest in that product.

c. Economic Situation – The economic situation of an individual affects the level of income at its disposal, savings and expenditure on particular groups of products.
d. Lifestyle – The individual's way of life, indicated by his interests, activities and expressed opinions and values.

4. Psychological Factors
    a. Motivation – This is the factor that directs the individual to fulfill his or her needs. The needs can take different forms: Biological, such as hunger, thirst or discomfort, or psychological, resulting from the need for recognition, affiliation or acceptance. Abraham Maslow's pyramid explains the hierarchical order that directs the individual to satisfy successive needs. This means that a person who has the need to satisfy hunger or thirst will not focus on the realization of personal development, but will focus on what is most important to that person at the moment. While the needs of the lower levels will be satisfied, the person can focus on the realization of the higher levels of the pyramid.

Figure 3 Maslow's needs hierarchy

Source: https://www.thoughtco.com/maslows-hierarchy-of-needs-4582571

b. Perception – The process in which people find, organize information in such a way that it takes on a meaningful perception of the world for them.
c. Learning - means a change in an individual's behavior along with the experience he/she has gained. Learning can take the form of positive or negative experiences that an individual has had in the past with a certain product.
d. Belief and Attitude – Describes what a given person thinks about a certain topic, what feelings drive that person, and what is his/her attitude towards given objects or concepts. Belief & Attitude is something that people gain through learning and perception, which ultimately influences their consumer decisions (Kotler & Armstrong, 2016).

Out of the 20 attributes available in the data set, 4 can be listed as describing a person. These are:

- Age
- Job
- Martial status
- Education

From the customer's point of view, a bank deposit is not a primary product. According to the needs indicated in the Maslow Pyramid, the deposit can be qualified between the second and the third level, allowing to increase the security for the future for oneself or the immediate family. For this reason, it can be expected that its target group will be consumers with family and most likely children as well. A bank deposit is also a product that does not bring immediate benefits, its value increases over time, while the available income is invested in the account. For this reason it is to be expected that customers interested in purchasing a bank deposit may be people of at least middle age, whose work provides sufficient earnings to cover basic expenses, and who decide to keep their savings for the future.

The correlation between the level of education and income per capita was shown in the 2006 study by Andres Redriguez-Pose and Vassilis Tselios, showing a positive correlation of almost 0.8 (Pose & Vassilis, 2006). This implies that a social group with a higher level of education should be more interested in acquiring bank deposits.

3.2.3 Macro economical factors

The customer's decision making is strongly determined by the environment. Although societies show varying degrees of differentiation, changes in the macroeconomic environment are nevertheless felt by a large percentage of societies. By characterizing a bank deposit as a luxury commodity, it may be concluded that unfavorable changes in the macroeconomic environment may result in less interest in purchasing this product. This section presents the impact of particular indicators on the sale of bank deposits.

Out of the 20 attributes in the dataset, 5 of them can be classified as macroeconomic variables. These are:

- Euribor 3 month interest rate
- Consumer Price Index (CPI)
- Consumer Confidence Index
- Employment Variation Rate
- Total number of employees

3 month Euribor interest rate is the rate at which wholesale funds in euro could be obtained by credit institutions in the EU and EFTA countries in the unsecured money market. EURIBOR is a critical interest rate benchmark authorized under the EU Benchmarks Regulation (BMR) (EMMI, 2019). EURIBOR, as a determinant of short-term interbank deposits, also has an impact on the end interest rate of deposits offered to customers.

The impact of interest rates on bank deposits can be explained by the classical interest rate theory. This theory assumes that when the supply of savings exceeds the demand for investment, the interest rate on savings will decline, discouraging savings on the one hand and encouraging

investments on the other. Similarly, if investment demand exceeds savings, the interest rate on savings increases to discourage investment and encourage savings (Dornbusch, 2011). A positive correlation between the increase in the interest rate and the number of deposits was also demonstrated in Fisnik Morina & Rufi Osmani's research on the impact of macroeconomic factors on the level of deposits in the banking sector. (Morina & Osmani, 2019).

Inflation measured by consumer price index (CPI) is defined as the change in the prices of a basket of goods and services that are typically purchased by specific groups of households. Inflation is measured in terms of the annual growth rate and in index, 2015 base year with a breakdown for food, energy and total excluding food and energy. Inflation measures the erosion of living standards. A consumer price index is estimated as a series of summary measures of the period-to-period proportional change in the prices of a fixed set of consumer goods and services of constant quantity and characteristics, acquired, used or paid for by the reference population. Each summary measure is constructed as a weighted average of a large number of elementary aggregate indices. Each of the elementary aggregate indices is estimated using a sample of prices for a defined set of goods and services obtained in, or by residents of, a specific region from a given set of outlets or other sources of consumption goods and services. (OECD, 2019)

The impact of inflation on the level of bank deposits may be similar to the one observed between the interest rate and the number of deposits. As the inflation rate is a factor that strongly influences the upward or downward movement in interest rates, it can be assumed that these variables have a strong positive relationship. The influence of inflation on interest rate can be best explained by *Fisher's Effect*. Fisher reached the conclusion that the expected rate of inflation and interest rates are in proportion to each other. The expected high inflation rate means higher interest rates, and vice versa. (Morina & Osmani, 2019) However, while the relationship between the interest rate and the level of bank deposits has been confirmed by Fisnik Morina & Rufi Osmani's research, the relationship between inflation and the level of deposits has been rejected. (Morina & Osmani, 2019). On the other hand, the authors of Altin Turhani & Dr. Hysen Hoda unexpectedly confirmed the negative correlation between the level of deposits and the interest rates. (Turhani & Hoda, 2016)

Consumer confidence indicator provides an indication of future developments of households' consumption and saving, based upon answers regarding their expected financial situation, their sentiment about the general economic situation, unemployment and capability of savings. An indicator above 100 signals a boost in the consumers' confidence towards the future economic situation, as a consequence of which they are less prone to save, and more inclined to spend money on major purchases in the next 12 months. Values below 100 indicate a pessimistic attitude towards future developments in the economy, possibly resulting in a tendency to save more and consume less. (OECD, 2019)

The value of this indicator above 100 may indicate positive consumer feelings about investments, which may result in a positive relationship with opening bank deposits. A value below 100 may be characterized by a negative impact on bank deposits' subscription.

Employment rates are defined as a measure of the extent to which available labor resources (people available to work) are being used. They are calculated as the ratio of the employed to the working age population. Employed people are those aged 15 or over who report that they have worked in gainful employment for at least one hour in the previous week or who had a job but were absent from work during the reference week. The working age population refers to people aged 15 to 64 (OECD, 2019). Variation of employment rate is defined as quarterly changes in the employment rate.

The impact of the employment rate on the number of bank deposits can be implicitly described by the relationship between unemployment and inflation. This relationship was observed by New Zealand economist William Phillips, named Phillips curve by his name. The Phillips curve states that inflation and unemployment have an inverse relationship. Higher inflation is associated with lower unemployment and vice versa (Phillips, 1958). Taking advantage of this relationship and the knowledge that inflation has a positive impact on the amount of bank deposits, it can be observed that the growing unemployment rate will have a negative impact on the amount of bank deposits.

The researchers Altin Turhani & Dr. Hysen Hoda came up with interesting conclusions, observing a positive relationship between unemployment and bank deposits in the period 2005-2014. (Turhani & Hoda, 2016)

The total number of employees represents the total number of workers employed in the Portuguese economy. When presenting the total value, it gives a narrower view than the relative unemployment rate, however, the effect of this variable on the level of bank deposits should be strongly correlated with the effect presented by the unemployment rate.

3.2.4 Client and campaign related factors

The next group of attributes present in the Portuguese bank's dataset are those related to the customer's credit history and those related to the ongoing sales campaign. They provide details about the customer's financial history in the current bank and how the customer was behaving while conducting a former marketing campaign. It also comprises information about the current marketing campaign.

According to P. Kotler and G. Armstrong, the purchasing process starts well before the actual point of purchase and takes significantly longer. The consumer decision-making process consists of five successive steps:

- Need recognition – First stage of purchasing. At this point the consumer spots a problem or need that he would like to satisfy.
- Information search – The moment when the client is driven by a motive to discover more information about products.
- Evaluation of alternatives – The point at which the consumer uses the information obtained in order to assess the various possibilities and select the most advantageous option for himself.
- Purchase decision – The actual time of purchase.
- Post purchase behavior – The stage in which the consumer takes additional actions after the purchase according to their own satisfaction or disappointment. (Kotler & Armstrong, 2016)

By knowing the cost of carrying out the campaign, information about the number of contacts made can help to assess the profitability of the entire campaign. Thanks to this, the bank is able to assess at what number of contacts the customer stops being profitable or even build a model

that selects the features needed to determine what kind of customers make decisions below the set efficiency threshold.

3.2.5 Data-driven approach of Data Analysis

The way of building the model that will be used in the paper can be specified as a data-driven approach. This implies an approach in which model objectives, patterns and relationships will not be based on assumptions made by the analyst, but will be discovered in the data. This method is beneficial as it allows to discover what really happens among the data and what drives the target variable. This methods is gaining from taking the practical approach of predictive analytics.

According to Dean Abbott the difference in perception and significance of the meaning of analysis between statistics and predictive analytics falls in its application. While the data analysis in statistics aims to confirm or deny a formerly assumed hypothesis, predictive analytics assumes the data analysis as a tool to find relationships and patterns within the data. This means focusing attention and acquiring information through what is " said " by the data.

> "*In spite of the similarities between statistics and analytics, there is a difference in mindset that results in differences in how analyses are conducted. Statistics is often used to perform confirmatory analysis where a hypothesis about a relationship between inputs and an output is made, and the purpose of the analysis is to confirm or deny the relationship and quantify the degree of that confirmation or denial."* (Abbott, 2014, p. 11)

To benefit from the information gathered through the data analysis, it is necessary to consider for what purposes it can be used. In the event of a sales campaign, a single telephone contact with a customer may be considered as a unit of analysis. Increasing the effectiveness of each telephone contact would allow to improve the results of the campaign by increasing the value of sales or achieving a similar result with less use of resources. Increasing the effectiveness of contact can be achieved through appropriate selection of the customer who will be contacted.

The authors Philip Kotler & Gary Armstrong defined the Customer-Driven Marketing process in two subsequent phases:

I. Market Segmentation - It's a process of dividing a market into smaller segments of buyers with distinct needs, characteristics, or behaviors that might require separate marketing strategies or mixes.
II. Target Marketing – It is a process of addressing the company's activities to specific market segments, which may seem to be the most favorable, through reaching customers who are most interested in the company's products or the most exposed to particular products.

As a result of market segmentation and following target marketing, it is possible to address the content of a marketing campaign to the group of customers who are characterized by the highest potential profit, and consequently to optimize the performance of the sales campaign. (Kotler & Armstrong, 2016, p. 223)

It is unlikely to expect that every time a telemarketer gets a list of potential contacts, he will take the customer' attributes, put them into model, calculate the likelihood of each of them buying a product by himself and then select those who will prove to be the most noteworthy. This is of course possible, but nevertheless labor-intensive. Decision supporting system (DSS) can come in handy. Author Vicki Sauter defines decision supporting system as follows: "*Decision support systems are computer-based systems that bring together information from a variety of sources, assist in the organization and analysis of information, and facilitate the evaluation of assumptions underlying the use of specific models. In other words, these systems allow decision makers to access relevant data across the organization as they need it to make choices among alternatives*". (Sauter, 2010, p. 5) By using an appropriate DDS, which could automatically predict the probability of selling a bank deposit using a previously created predictive model, telemarketer could be provided with the group of potential customers who has scored the highest ranks and thus, contacted in the first term.

3.3 Techniques and methods of data processing

According to the CRISP-DM model, the phases following the Business understanding phase focus on working with data. To build a model that will be able to address the needs of the business and bring the desired value it is necessary to prepare the data for modelling first. In order to effectively improve the quality of data and bring it to a useful form it is necessary to understand the data, carry out data cleaning and deal with missing data. Data in a raw form can create an obstacle for certain algorithms, causing their accuracy and usability to fall.

3.3.1 Data understanding

As mentioned in CRISP-DM, the next process after business understanding is data understanding. It is the process of having a first look inside the data, evaluating its quality and providing a reference point for the subsequent modelling steps. D. Abbott defines the following tasks that should take place in the data understanding phase:

- Determination of descriptive statistics for key variables that are subsequently used in the modelling process.
- Discovering and prioritizing the problems that arise in the process of data observation, including data with erroneous values, missing data as well as incorrect data distribution or outliers.
- Visualize data to discover a deeper insight into the data. It allows to see parameters that are not evident using descriptive statistics. (Abbott, 2014)

One of the ways to understand the data is to use descriptive statistics. Descriptive statistics is therefore a data analysis that helps to find trends and patterns within the accumulated data. It is a method of describing data in such a way as to depict what information it contains. Descriptive statistics contains two groups of measures. These are *measures of central tendency* and *measures of dispersion.*

Central tendency measures is a group of statistics describing a data set in a centralized manner, providing a single value corresponding to the whole set. The measures of central tendency are:

- Mean

- Median
- Mode
- Skewness
- Kurtosis

Among the central tendency measures, skewness and kurtosis are particularly important from the perspective of predictive analytics. Skewness is a measure of distributional balance. A normal distribution has a skewness of 0. Positive skewness means that the distribution has a tail on the right side of the central values, while negative skewness refers to the tail on the left side of the main body of the distribution. The effect of skewness on the distribution is shown in Figure 4. This also shows how skewness affects the other values of the central tendency. Both the median and the mean are shifted to the right in case of a positive bias and to the left in case of a negative bias respectively. The measurements of the skewness allow for the detection of outliers, i.e. results that differ significantly from other observations. A positive bias allows to identify values significantly larger than the main part of the distribution, while a negative bias indicates values significantly smaller.

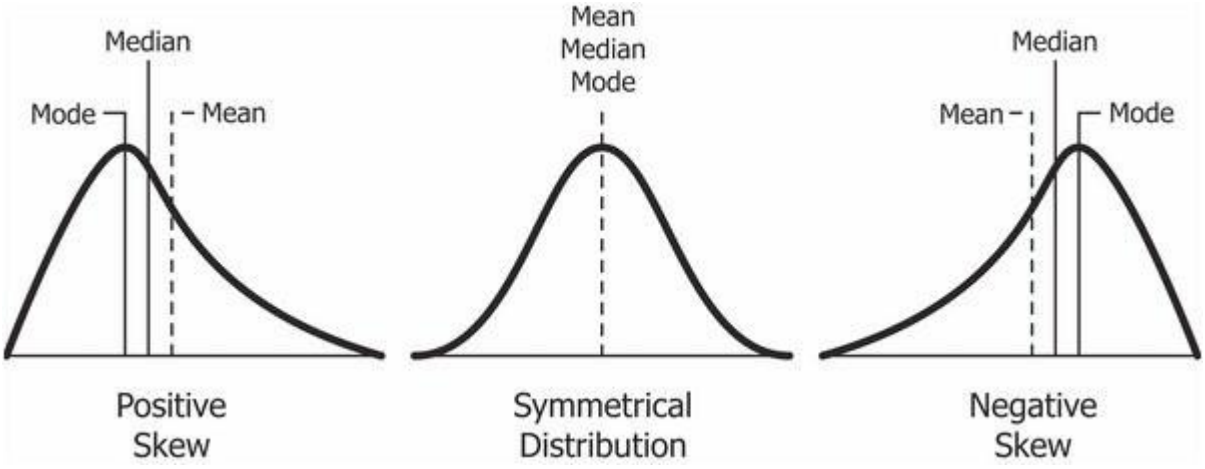

Figure 4 Skewness of distribution

Source:https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa?gi=8467a4b8c6ea

Kurtosis is a measure of how thicker or fatter the tails of distribution are in comparison with normal distribution. A kurtosis of 3 is referred to as Mesokurtic and means normal distribution. Values above 3 are referred to as platykurtic and denote concentrations of observations around

the tails. A Kurtosis value below 3 is referred to as leptokurtic and denotes a concentration of the distribution around the center. Figure 5 shows distributions with different Kurtosis is not usually used in predictive analytics for tasks except for providing information about the distribution. However, it helps to determine whether the analyzed distribution of data is similar to normal, uniform or neither.

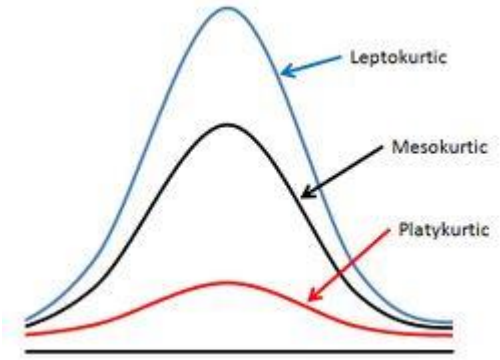

Figure 5 Kurtosis

Source: https://www.bogleheads.org/wiki/Excess_kurtosis

Measures of dispersion supplement the information missing within central tendency measures, giving a complete picture of descriptive statistics. They refer to the distribution of data around the previously mentioned central values. They help to determine the degree of dispersion within the distribution. The dispersion measures are:

- Variance & Standard deviation
- Min
- Max
- Percentiles
- Range
- Quantile
- Correlation

Many of the dispersion measures are used for the first insight into the distribution of data and identification of outliers. The standard deviation is a measure that indicates how much the average observation differs from the mean. Knowing the assumption of normal distribution is

that 99.7% of observations should be located within 3 standard deviations from the mean it can be assumed that observations beyond 3 standard deviations can be considered as unexpected values and qualify them as outliers.

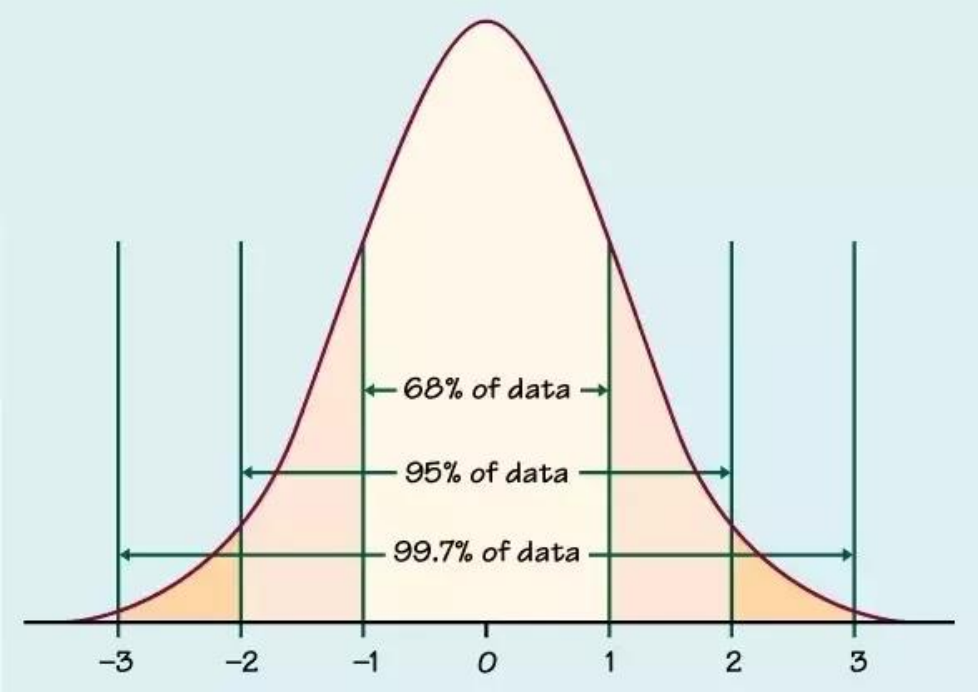

Figure 6 Normal distribution

Source: *https://www.quora.com/Can-artificial-intelligence-help-in-eradicating-poverty*

Measures such as min, max, percentiles, quantiles allow to determine what values are represented by the distribution at particular points and what is the numerical range covered by the analyzed distribution. They are also called rank-ordered statistics because they refer to values in sequence from minimum to maximum. These measures are considered to be robust statistics because their use is not limited by the type of distribution and they fully reflect the characteristics of the distribution. Percentile are distribution values taken at individual percentages of the scale from minimum to maximum. Quantiles are values taken at 25, 50 and 75 percent respectively. A common method of identifying unusual values is to use a value equal to 1.5 times inter-quantile-rate. For this purpose, the 1.5 IQR distance from the 1^{st} quantile and 3^{rd} quantile are defined as the break-points above which observations are considered as outliers.

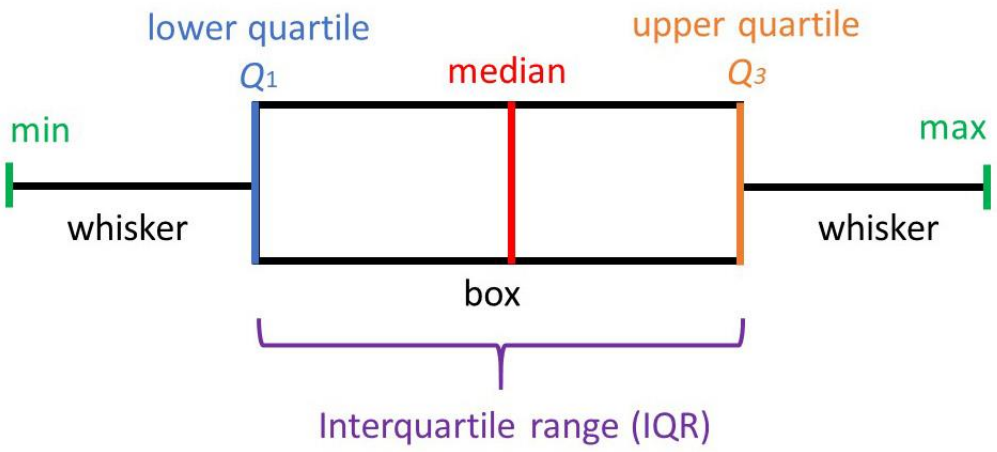

Figure 7 Interquartile range

Source: https://www.simplypsychology.org/boxplots.html

Correlation measures the numerical relationships between variables. It is a useful measure to determine which pairs of variables affect each other. The discovery of variables with high correlation indicates that the variables explain the same events, so their presence during modelling may turn out to be redundant. (Zhang, 2017)

The above-mentioned measures and methods referred to the measurement and analysis of the numerical variables distribution. Categorical variables are measured by the number of observations at specific levels, providing information on what values are most common in a data set. For this purpose, measures of the number of occurrences and the percentage share of a given value in the entire population are used. This allows for a first view on the data set. (Sarkar, 2020)

Another way to view data is to visualize it. Visualization also involves aggregating data, but it does so in a way that allows to discover additional information and context. Certain features that are difficult to convey in a numerical way can easily be presented visually. An example of such data is the density and distribution of data. By presenting the distribution of data graphically, it is possible to discover the occurrence of outliers in a much clearer way. This also makes it possible to identify whether the data has a normal, uniform or perhaps a completely different distribution.

One of the ways of presenting one-dimensional data is a histogram. Usually on the X axis they have values of variables and on the Y axis the frequency of their occurrence. A sample histogram is presented in figure 8.

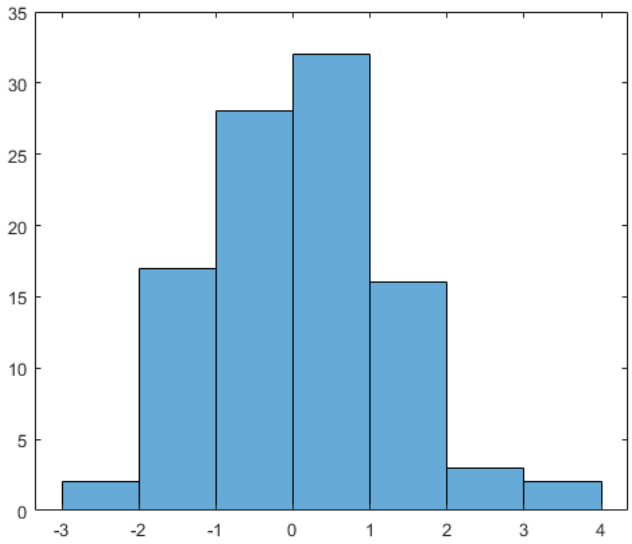

Figure 8 Histogram

Source :https://www.liberaldictionary.com/histogram/

Histogram is an effective way to present the distribution of one variable. To present two or more dimensional data scatter-plots can be used. Scatterplots are a common way of showing relations between variables, if their correlation is significant then the data will be arranged in the shape of a cigar Figure 9 presents 2 and 3 dimensional scatterplots. (Abbott, 2014)

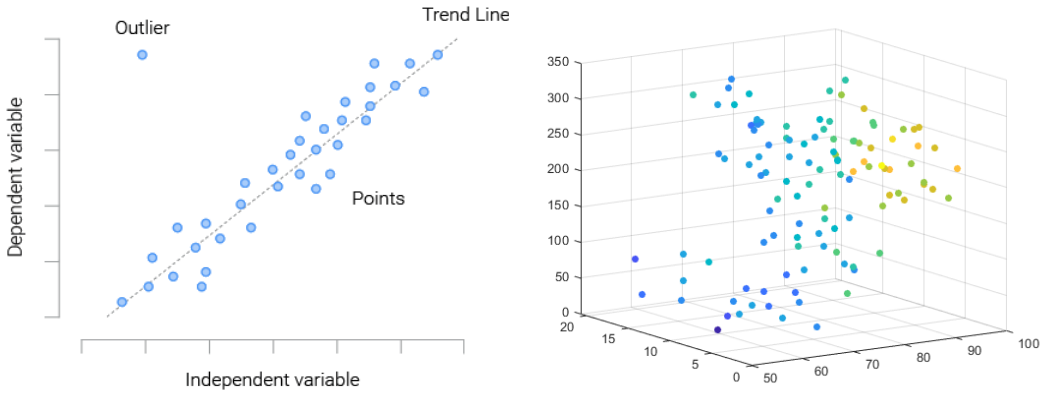

Figure 9 Scatterplots

Source: https://www.learnbyexample.org/r-scatter-plot-base-graph/
https://se.mathworks.com/help/matlabmobile_android/ug/creating-3d-scatter-plot.html

The reason for the importance of the data insight is the need to catch the flaws in the distribution. Faults that may occur in a data set can be classified into the following categories:

- Missing data
- Misclassified data
- Outliers
- Non-sense values

Identifying and subsequently cleaning the data is a key element in creating predictive analytics models. Due to the fact that numerical algorithms assume a distribution similar to normal or at least uniform distribution and are very sensitive to outliers or missing data, it is very important to identify and subsequently manage these occurrences. Building numerical algorithms based on the distances between numbers can give pointless models without proper data preparation.. (Larose & Larose, 2015)

3.3.2 Data Cleaning and preparation

Following the CRISP-DM methodology, the phase following data understanding is data preparation. In the data understanding phase, using descriptive statistics and data visualizations, flaws and problems should be already highlighted, thus the data preparation phase should fix them. Given information and a first insight into the distribution, data quality and presence of outliers, one should now deal with cleaning, preparing and transforming the data into a form which is suitable for usage of modelling algorithms.

The first stage of data preparation is data cleaning. It consists of fixing previously diagnosed problems inside the data, including miscoded values, missing values and outliers. Repairing these problems is key to building robust predictive models. The problems to be confronted when cleaning variables are as follows:

- Incorrect values
- Outliers

- Missing values

In order to determine the validity of data, it is necessary to define which values are qualified as correct. Incorrectly encoded categorical data may have a low frequency of occurrence, but in order to clearly define what values are expected within a variable, it is necessary to consult with the domain expert who is able to determine at what point the data takes unexpected values. Numerical variables containing incorrect data will manifest themselves as outliers. In case the data cannot be decoded, it should be treated as missing, except in the case of high numbers. Then they should be left for later interpretation.

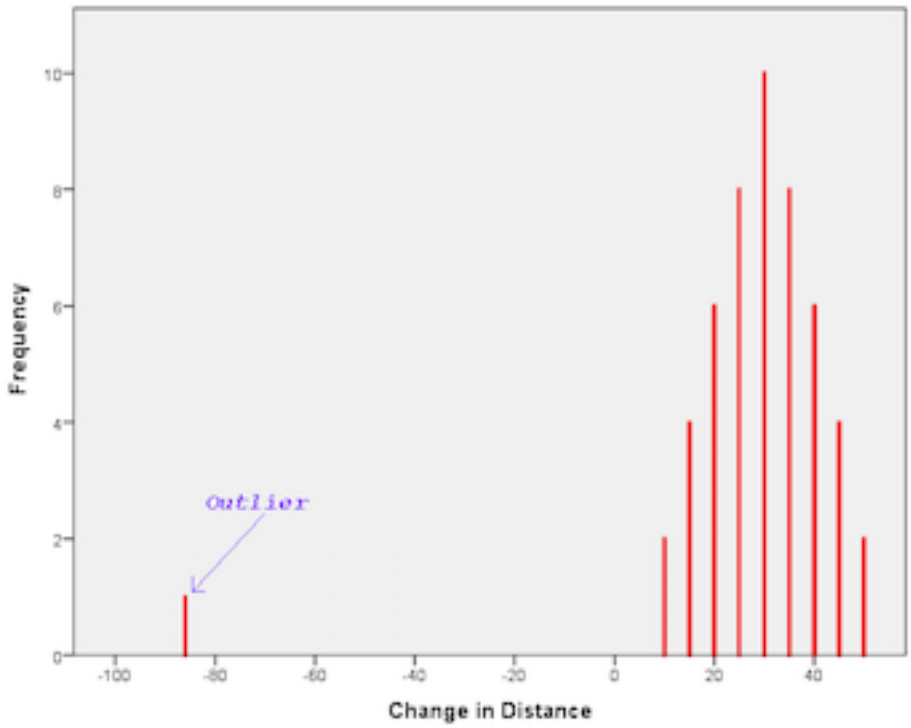

Figure 10 outlier

Source: https://study.com/academy/lesson/outlier-in-statistics-definition-lesson-quiz.html

Outliers, i.e. unexpected values inside the variables, usually significantly separated from the main part of the distribution can significantly affect algorithms which use numerical distances. Example of outlier is visible on figure 10. Outliers can be handled in the following way.

1. Remove outliers from the data
2. Separate outliers and create a new model just for them

3. Transform outliers, that they are no longer outliers
4. Bin the data
5. Leave outliers in the data

The way of dealing with outliers depends largely on the needs of the analyzed data and the chosen modelling method.

Missing values are considered to be the most problematic group of data fixing problems. In order to correct missing values, at any possible case, one should try to insert a values into them.

In case this is not possible to do so, and the analyzed project allows it, listwise or column deletion is possible. This means, in the first case, deleting observations where missing values are present or, in the case of column deletion, removing the whole attribute if missing data are present. This way, however, carries a significant risk of losing the information contained in the deleted data. Nevertheless, if possible, the best option is to replace missing value with imputation. It is possible to insert values from the following sources:

1. Constant inputation,
2. Mean / median inputation,
3. Inputation from distribution,
4. Inputation from distirbution

If missing values are replaced by a fixed number, this may cause a distortion of the distribution, if it occurs frequently, so it may be preferable to replace missing values with values from the distribution or from the model. Additionally, inserting data from the model can be used for categorical variables.

Sometimes missing data can contain information in itself. This allows them to be included in dummy variables and then used to build the model. (Abbott, 2014)

Feature creation refers to variables added to a dataset that have been created from one or more variables already within the data.

Fixing skew is one of the steps that can be performed at a feature creation stage. When the skewed distribution is being fixed, the variables are changed by the functions that affect the

extreme values in the tails to the greatest extent. The most desirable effect of a skewed distribution is to bring the distribution to a state similar to normal distribution. Table 1 shows how to correct different types of skews.

Problem	**Transform**
Positive skew	$log10(1+x), \frac{1}{x}, \sqrt{x}$
Negative skew	$x^n, -log10(1+abs(x))$
Big tails on both sides	$sign(x)log10(1+abs(x))$

Table 1 Skewness fixing methods

Of the methods listed above, logarithmic transformations are the most commonly performed transformations to reduce the bias. 1 added to the logarithmic value prevents the problem of zero transformation. (Abbott, 2014)

A method that solves the problems of abnormal variable distribution is to create a categorical variable using a binning technique. This means defining the boundaries of bins within a given variable. Then, in order to transform a categorical variable into a numerical one, each level of the variable can be transformed into a corresponding binary variable. Figure 11 shows a binning distribution. (Finlay, 2014)

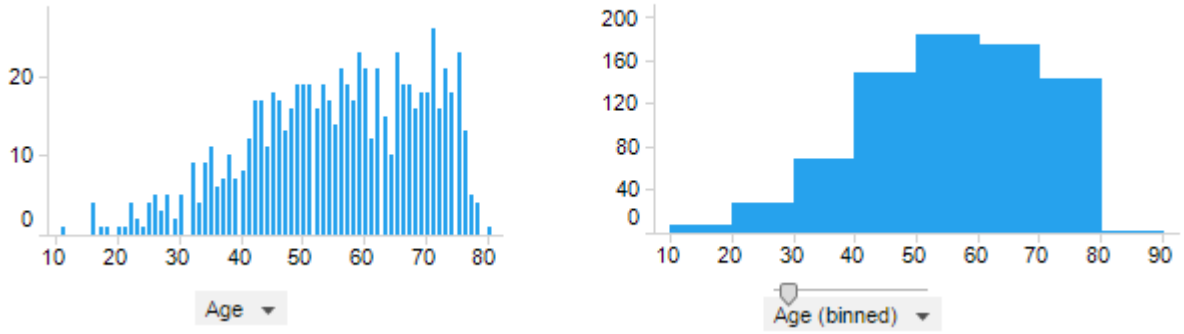

Figure 11 binned distribution

Source:https://docs.tibco.com/pub/sfire-bauthor/7.5.0/doc/html/en-US/GUID-2CA43500-C5CE-40B4-88BD-0E22438F03FB.html

The method used to reduce the distances between the data is normalization. As a result, the numerical variables take the expected range. In this way the influence of large sizes is reduced.

This is especially important for numerical algorithms based on linear distances. Table 2 shows a list of the most popular techniques used to scale variables.

Method	Formula	Expected range
Magnitude scaling	$x = \frac{x}{\max \lvert x \rvert}$	$[-1, 1]$
Sigmoid	$x = \frac{1}{(1 + e\left(-\frac{x}{c}\right))}$	$[0, 1]$
Min-max normalization	$x = \frac{(x - x_{min})}{(x_{max} - x_{min})}$	$[0, 1]$
Z-score	$x = \frac{(x - x_{mean})}{x_{std}}$	$[-3, 3]$
Rank binning	$\frac{(100 * rank\ order)}{\#\ of\ records}$	$[0, 100]$

Table 2 Normalization methods

A good practice when scaling variables is also to keep one scale. While some of the variables will contain, for example, values in the range [0, 1] and the rest will remain in the nominal scale, the weights of individual parameters may be distorted. (Abbott, 2014)

As some numerical algorithms are unable to accept variables other than numerical, the categorical variables must be converted accordingly. For this purpose, a variable transformation is used which consists of listing all categories of a given attribute and then encoding the corresponding value with a binary variable. This carries the risk of introducing too many zeros, while the categorical variables are numerous, so it is common practice to code 0,3 as none and 0,7 as presence.

Having a set of prepared data, the next step is to verify which variables will be needed in the next modelling steps. Keeping the variables that do not bring additional information to the model is unnecessary. Variables that should be removed are primarily unary variable and nearly unary. Unary means that the variable has one value for each observation. In the case of nearly unary variables, the matter is not entirely obvious, because in some situations it is exactly that other value occurring in only 0,05% of the data may be crucial. However, if this is not relevant from the model point of view, then it can be removed.

D. Larose & T.Larose also advise to be careful when removing variables that have 90% missing values or strong correlation with another attribute. They believe that one should be particularly

careful because in these situations the information contained in 90% missing values, for example, may be associated with a systematic but unobserved phenomenon. If the records are really duplicates, they should be removed from the data set because they affect the bias of the data set. (Larose & Larose, 2015)

Feature selection is also referred to as Attribute selection or Variable selection and is part of Feature Engineering. It is the process to select a subset of most relevant attributes or features in the data set for predictive modeling.

As inappropriate features may adversely affect the performance of the model, it is worth considering feature selection before building the model. This can have a positive effect on the speed of model training, while instead of a few hundred or even several thousand features only the most appropriate ones will be selected. This will also increase the transparency of the model and help reduce over-fitting. (Khandelwal, 2020)

Amongst the techniques of selection features, there are two main types. These are the filter method and the wrapper method.

Filter method ranks each feature on the basis of statistical tests and selects those that have satisfactory value. Among the filter measures we can mention:

- Variance
- Chi-square test
- Anova
- Correlation coefficient
- Information gain

Wrapper methods search among a set of attributes that are best suited for input to the model. These methods use comparisons with previous versions of the model to determine whether a new feature should be added or removed. The following methods are used for this purpose:

- Forward selection – starts with null features and then adds more by measuring accuracy gains. If a variable improves the quality of the model, then it is added.
- Backward selection – starts with all variables and gradually removes them to compare if the overall quality of the model has been improved.

- Stepwise selection – In this case the model also starts with all variables and constantly evaluate their progression, but in contrast to backward selection at all times also tries to add other features. (Kumar, 2011)

Overfitting is a major threat to the models. It consists in the fact that the model teaches very precisely the patterns in the data set used for training to such an extent that its subsequent performance on another data set is unsatisfactory. To avoid this phenomenon one should consider evaluating the model on the data that the algorithm did not see when building the model. In this way we can create a division into training and testing datasets. To this aim, D. Abbot presents the following sampling data process:

1. Build a model on training data
2. Assess the model on testing data
3. Change settings, parameters or inputs to the model
4. Re-build model on training data
5. Assess the new model on testing data
6. Change parameters and repeat the process

Sometimes algorithms are also able to learn to test the data set, so in order to prevent this, a third data set should also be used to validate parameters. Nevertheless, each subset should contain sufficient number of records to be fully representative. The rule of thumb is that for each attribute there should be 10-20 observations in the dataset. However, more complex problems may require even 100 observations.

In situations where a data set contains so few records that further portioning is not possible, two of the sampling methods can be used to help split observations. These are K-folds cross validation and bootstrap sampling.

K-folds cross validation is a sampling method usually used in a small dataset. K corresponds to how many sub-sets of data are used for modelling. Cross validation consists of dividing the entire data set into k-elements followed by using k-1 for training and 1 subset for testing. Then the roles inside the dataset change and so each subset is used as a validation set. Figure 12 illustrates the cross validation scheme.

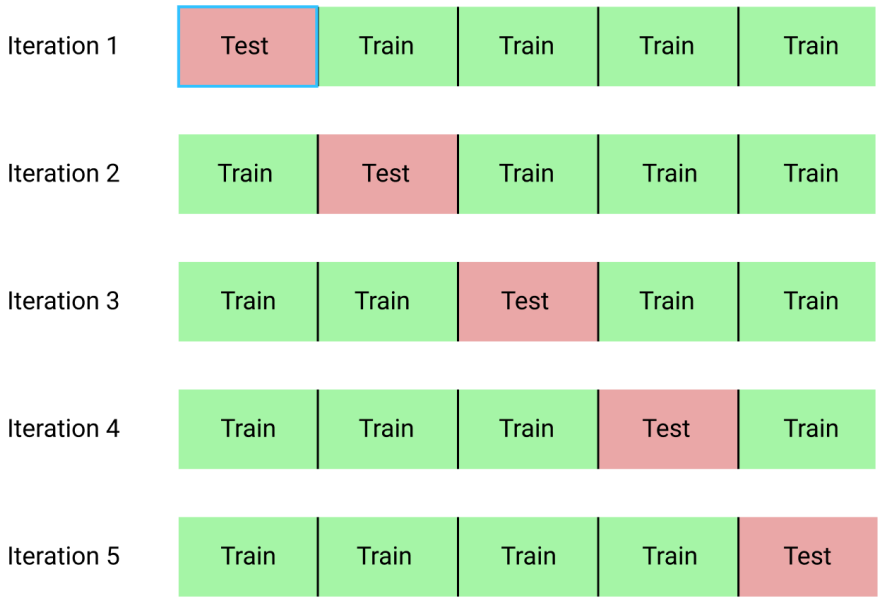

Figure 12 Cross validation

Source:https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

The second way to build quality models on small data sets is bootstrap sampling. This method uses the creation of sample sets by randomly taking observations from the original dataset a certain number of times. Then the created number of k-quantities of samples is treated the same way as in the case of cross-validation. Bootstrap sampling allows to repeat data, so many observations will be duplicated, and some of them will not be inside certain samples. It is estimated that 1/e, or about 63% of the observations of the original data set will be found in samples created in this way. This allows to create different data sets for training and testing. The way Bootstrap sampling works is shown in Figure 13. (Abbott, 2014)

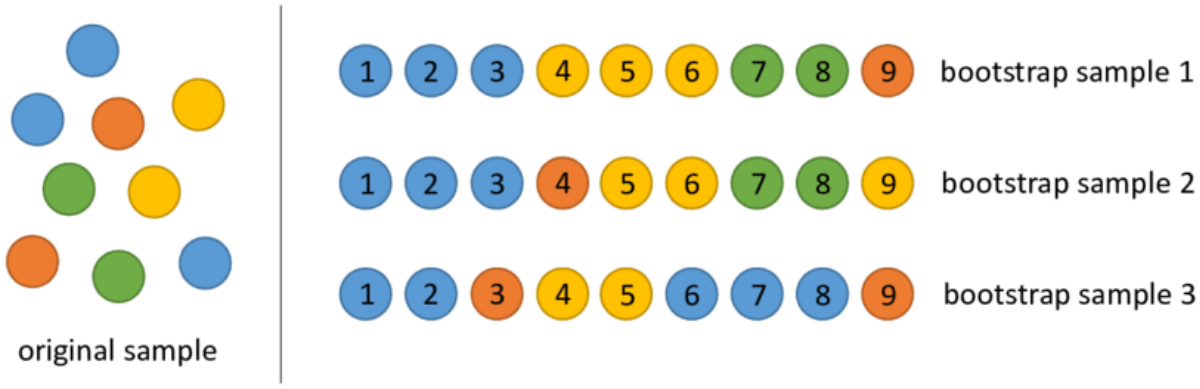

Figure 13 Bootstrap sampling

Source:https://www.researchgate.net/figure/An-example-of-bootstrap-sampling-Since-objects-are-subsampled-with-replacement-some_fig2_322179244

Using the techniques described above, data can be fed from a raw form, to one that enables the development of robust and accurate models. It should be taken into consideration that each problem has its own specification and this should not be considered as a rule, but rather as a guide and method.

3.3.3 Dimensional reduction

Data sets used in the data mining process frequently contain many millions of rows or thousands of variables. This means that the appearance of the problem of *multicollinearity*, which means variables so strongly correlated with each other that knowing the value of one of them you can predict the value of the other (Glauber, 1964) is very likely. As an example of such variables we can describe people' earnings and their net assets. Using variables that have such a strong influence on each other might lead to unreliable results.

Another issue that occurs in data sets containing a significant number of attributes is the Curse of Dimensionality phenomenon described by Richard E. Bellman. It means that when the dimensionality of data increases, the amount of the corresponding plane often increases exponentially in relation to the number of dimensions. For this reason, the amount of data required to cover the plane and build the model can also grow exponentially. Additionally, building a model of such complexity significantly increases the risk of *overfitting* (Bellman, 1957).

Data dimensioning can be reduced in two ways:

1. Feature Selection – Selecting variables which provide the most information.
2. Feature Extraction – Create new variables that will describe the current variables in the best possible way, extracting from them maximum information.

In the previous chapter we described the methods of selection of appropriate variables based on backward, forward and step-wise selection, therefore this chapter will describe feature extraction methods.

Principal Component Analysis (PCA) is one of the most popular method of Feature Extraction. This is a statistical activity that uses orthogonal transformation to transform a set of correlated variables into a set of linearly uncorrelated variables known as principal components. This transformation is achieved in such a way that the first principal component has the greatest possible variance and each subsequent principal component has the greatest possible variance under the condition that it remains perpendicular to the other principal components. An example of a two-dimensional transformation is shown in Figure 14. The result of these transformations is an uncorrelated orthogonal base set in a number equal to the number of variables of the initial data set. PCA is as well sensitive to relative scaling of the original variables. (Pearson, 1901)

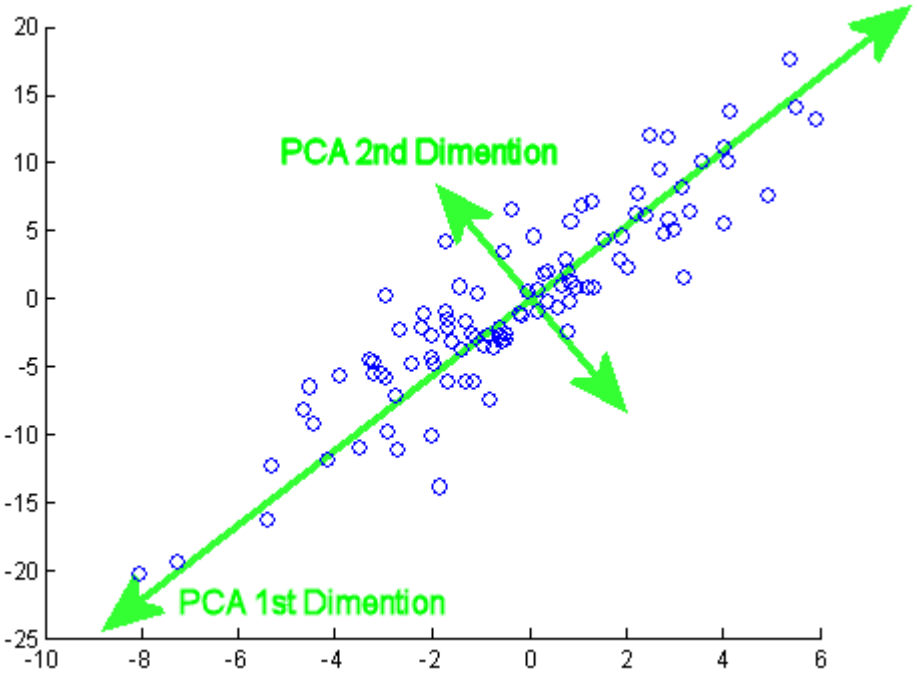

Figure 14 PCA algorithm

Source: http://weigend.com/files/teaching/stanford/2008/stanford2008.wikispaces.com/file/view/pca_example.gif

Thus, PCA is able to create a set of k linear components with m variables in such a way that a smaller number of k is able to contain almost as much variance as the entire set of m variables, thereby contributing to a reduction in dimensionality while still preserving the information transmitted. Each PC also contains Eigenvalue, which is the proportion of explained variance of all attributes. For example, PC1 with an eigenvalue of 3.5 means that the first PC explains exactly 'variations of 3.5 attributes' Eigenvalues are also used to identify the % variance explained.

Consideration should be given to what criterion should be taken into account when selecting how many principal components should be extracted. Daniel T. Larose and Chantal D. Larose propose the following methods to decide how many components should be selected:

- Eigenvalue criterion refers to the number of variable variances explained by each principal component. The most common threshold is the eigenvalue above 1.

- Proportion of variance explained criterion – The explained variant decreases with each of the subsequent principal components. Depending on the purpose of the principal component analysis and the field of science, the assumed value of the variance may vary, but usually takes a value of 90-95%. This means leaving enough PCs to reach this level of explained variance.
- Scree plot criterion – Scree plot is a graph that shows the % of cumulative variance explained by successive principal components. The intuitive rule is to find the moment after which each subsequent component added does not explain the high percentage of variance. The scree plot is shown in figure 15. In this case the % of the explained variation is graphically flattened out after 4 components.
- Minimum Communality Criterion – Communalities means the overall role of each of the initial variables within the PCA. For example, a variable with less communality than the others means that it contributes significantly less to the overall PCA. A very low communality value should indicate that the variable is not explained by the PCA. Communality values are calculated from the sum of the squares of the weights of the components for each variable. Communalities below 0.5 are considered low and if this value is not reached, subsequent PCAs should be extracted. (Larose & Larose, 2015)

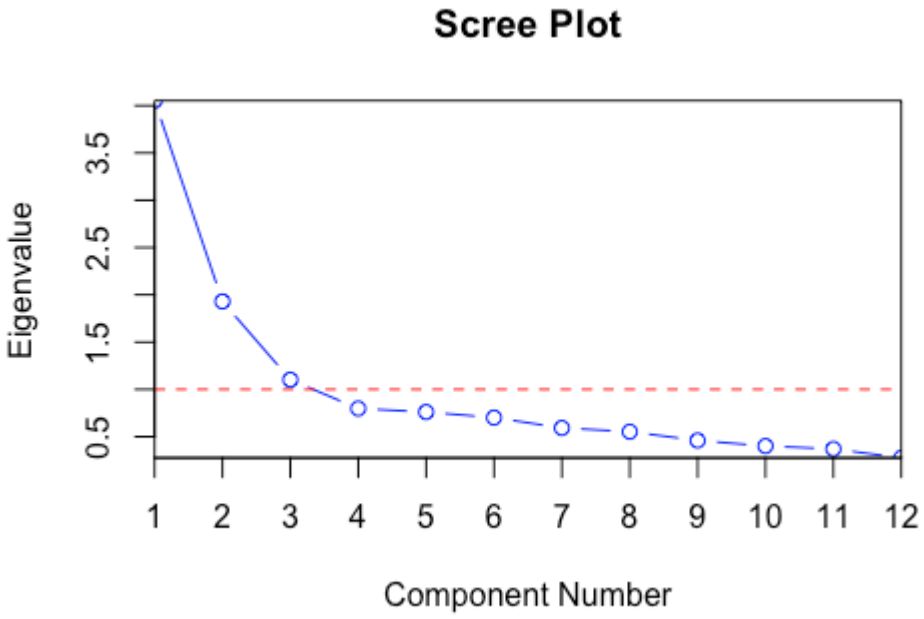

Figure 15 Scree plot

Source: https://en.wikipedia.org/wiki/Scree_plot#/media/File:Screeplotr.png

However, the PCA has its limitations. Due to its linear dependencies, it does not work for non-linear relationships between variables. It also assumes that data will have a normal or uniform distribution. While the use of PCA usually does not increase the predictive capabilities of models, it is also often practice to use PCA to verify which variables are most relevant within a dataset and then select the variables themselves. In addition, the use of PCA will not allow to limit the initial dataset, even if a satisfactory number of PCs are selected for use inside the model, the full number of initial variables will still be needed.

3.3.4 Descriptive modelling

The substate-based purpose of descriptive modelling is to provide more information about the input data. As well as detecting basic parameters with descriptive statistics, it also determines the relationship between the input data by assigning them to appropriate clusters. These are adequate techniques in situations where the target variable cannot be further specified or the sole purpose of the study is to find interdependencies within the data. (Falcidieno, Pienovi, & Spagnuolo, 2005)

The method used for descriptive modelling is a branch of statistical learning called *unsupervised learning*. It consists in discovering the best methods of data segmentation in such a way that the algorithm used is able to find the best possible way to detect the attributes that lead to the creation of *segments*. One of the tools of statistical learning used in unsupervised learning is clustering. The task of cluster analysis is to determine on the basis of values of individual attributes to which group this observation belongs. An example of cluster analysis can be market segregation of customers. During this analysis, individual attributes of customers are verified, such as home address, income, family size and shopping habits. A later goal is to create customer groups with characteristic features, called later clusters. The identification of clusters within customers may prove useful due to differences between the groups, which may affect for example the interest in company's product. (James, Witten, Hastie, & Tibshirani, 2017)

K-Means algorithm is an example of cluster analysis. The algorithm is a logical and computationally simple but very effective in creating data groups. It is an iterative algorithm that calculates the distance of individual points in space from the center of the cluster's and then assigns an appropriate label for observation. The iterative process of the algorithm is presented in the following steps:

1. Initialization – K number of clusters must be specified for creation inside the data set.
2. Step 1 – Each K cluster is assigned a randomly selected point called center of cluster.
3. Step 2 - The distance between the points of the plane and the center of the cluster is calculated.
4. Step 3 – Each observation is assigned a label indicating the nearest cluster center.
5. Step 4 – The average value of the cluster center is calculated on the basis of observations assigned to it. After, previous cluster center is being replaced by the new one.
6. Step 5 – Step 2 and 4 is repeated until the subsequent iterations do not change the value of the clusters and the observations do not change its labels.

An example of created clusters is shown in figure 16. Despite its usefulness in creating groups, the K-means algorithm also has its limitations. As it operates on linear distances, all input data should be converted into numbers. In addition, the data should have a distribution close to uniform or normal, and the values should be scaled by min-max or z-score transformation. (Abbott, 2014)

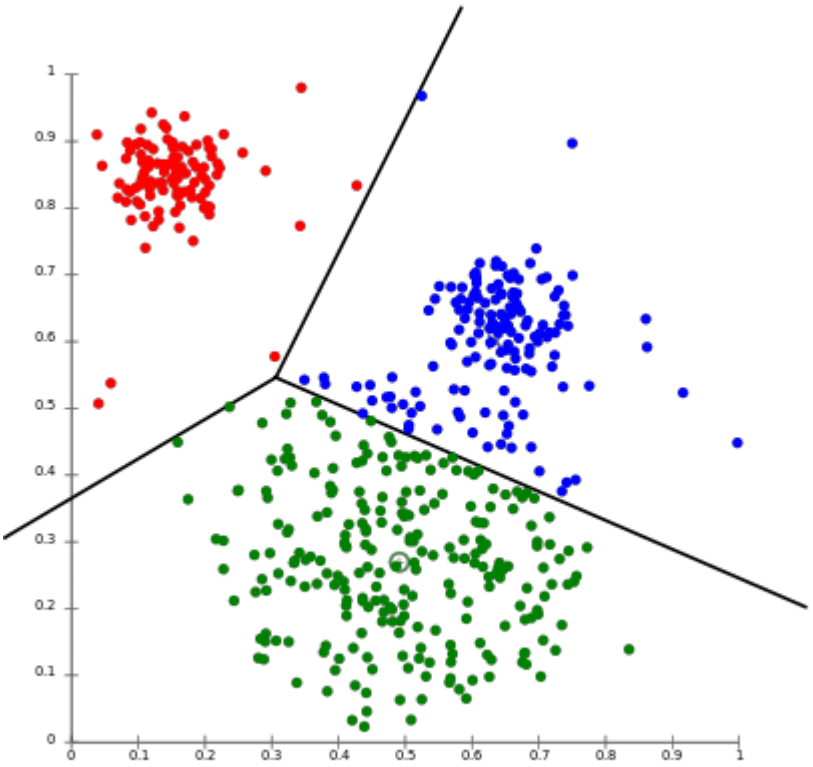

Figure 16 K-means

Source:https://aws.amazon.com/blogs/machine-learning/k-means-clustering-with-amazon-sagemaker/

3.4 Predictive modelling algorithms

Predictive modeling algorithms belong to the supervised learning group. This means that they try to find correlations between the input data and the target variable. The key objective of predictive models is to discover target variable values in the future based on the input data known today. The two main problems solved during supervised learning are classification and regression. The task of classification is to assign a class label available from a predefined list of possibilities. Often the classification is limited to a choice between two possibilities in cases such as whether the client has decided to buy a product or not. Here, the so-called binary classification is considered. In the problem of binary classification one class is often indicated

as positive and the other as negative. A positive class does not refer to a higher value, rather it refers to the presence of the object under investigation. The regression problem means predicting a previously undefined number. An example is an attempt to predict the earnings of a certain unit. (Müller & Guido, 2017)

As a result of the problem in the currently examined data set, the algorithms described below will address the classification problem or will be presented from this perspective. There are many robust algorithms that solve predictive problems and each of these algorithms has different versions. Among the most popular algorithms used for binary classification we can distinguish:

- Logistic regression
- Decision Tree
- Artificial Neutral Networks
- Support vector Machine

3.4.1 Logistic Regression

Logistic regression is a partially linear classification used for binary classification. This algorithm is used to describe the relationship between the input data set and the binary target variable. The occurrences are determined in logistic regression using the so-called Odds ratio. It can be interpreted as a proportion of the occurring events.

$$odds\ ratio = \frac{P(1)}{1 - P(1)} = \frac{P(1)}{P(0)}$$

The logistic regression parameters are established using maximum likelihood function. Just as the likelihood function is to show the degree of matching of the statistical model to the data sample, the maximum likelihood estimation uses the gradient descent to discover the parameters which have the best matching. With the regression parameters being determined, the odds ratio of logistic regression can be presented as follows:

$$Log\ odds\ ratio = w_0 + w_1 x X_1 + \cdots + w_n\ x\ X_n$$

The probability of an event, also known as a logistic curve. By plotting points on the plane, the shape of the function visible on figure 17 is created. Unlike linear regression, the logistic regression function takes a shape similar to the letter "s". The logistic curve takes values from the interval [0;1] usually the value of 0,5 can be interpreted as decision boundary between two classes. The probability of the event is expressed by the following function:

$$P(target = 1) = \frac{1}{1 + e^{-(w_0 + w_1 x X_1 + \cdots + w_n \, x \, X_n)}}$$

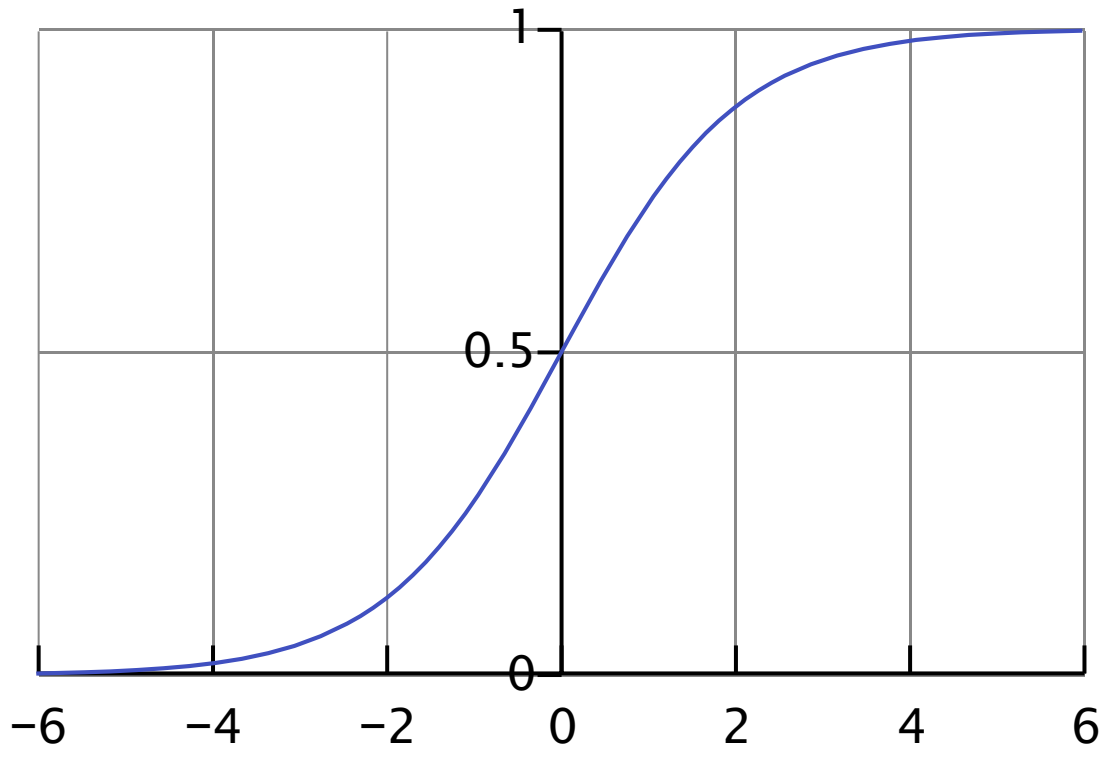

Figure 17 Logistic curve

Source: https://en.wikipedia.org/wiki/Sigmoid_function#/media/File:Logistic-curve.svg

Logistic regression is a numerical algorithm. It means that during data preparation, all categorical variables should be converted into numbers. This means the transformation of categorical variables with n levels into n-1 dummy variables. Additionally, logistic regression cannot handle missing values. If they occur, one of the methods of imputation should be used. (Larose & Larose, 2015)

3.4.2 Decision Tree

Another widely used supervised learning algorithm is the decision tree. They are used for both classification and regression purposes. The decision tree learns the series and hierarchy of if/else questions leading ultimately to a decision. This feature enables extremely clear interpretation of decision trees. An additional advantage of decision trees is their simplicity in implementation. Data used for modeling does not have to be converted into numbers, the algorithm can handle both numeric and categorical data. Decision tree does not assume parametricity, thus allowing it to work on distributions other than normal and uniform. An example of a decision tree is shown in figure 18.

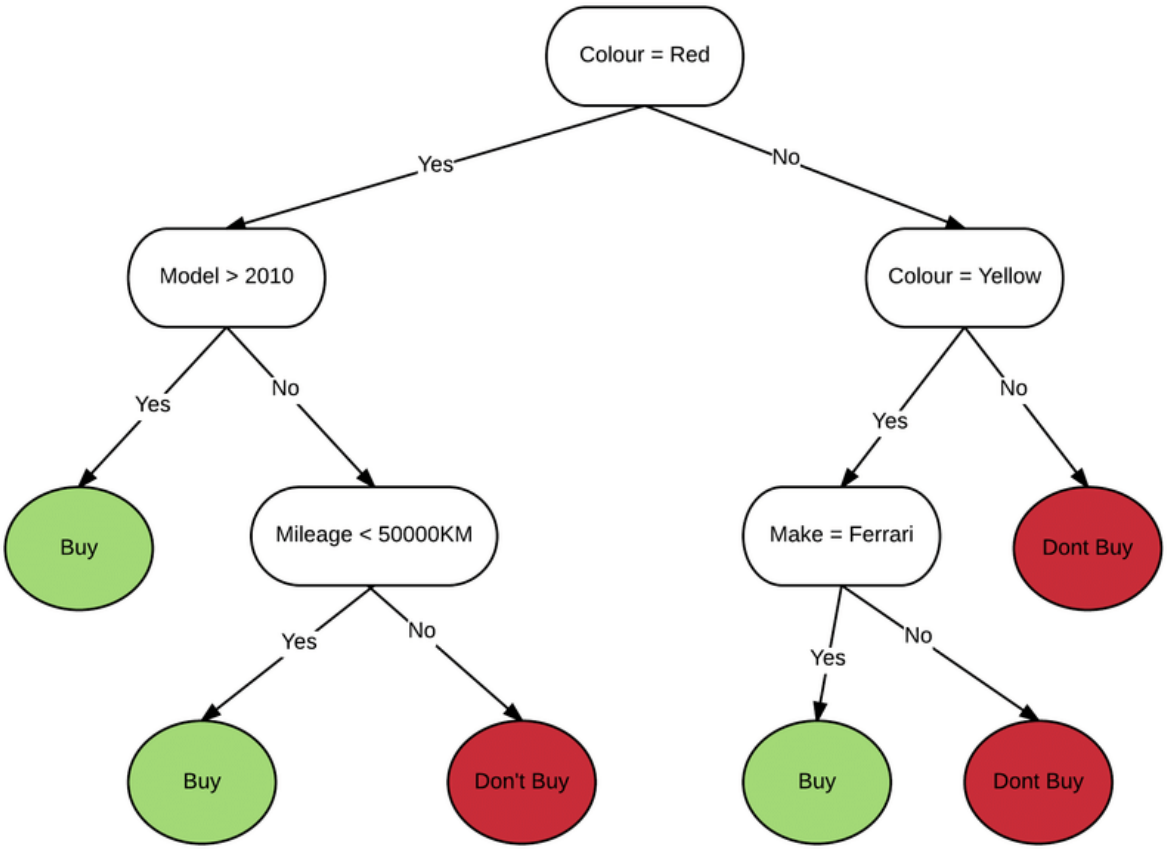

Figure 18 Decision Tree

Source:https://www.researchgate.net/figure/An-example-of-a-simple-decision-tree_fig2_311614501

The terminology of the decision tree is essentially similar to that of the actual tree. The difference is that the decision tree has roots at the top and leaves at the bottom. Split is a condition that divides data into two or more sub-groups. The first split inside a tree is called root split and is often considered to be the most significant inside the decision tree. Each subsequent subgroup created during the split is called a branch. The decision tree can be described as a recursive portioning algorithm. The way it works can be described as follows:

1. For each of the inputs, find the best possible way to split into two or more subgroups. From the division methods, choose the most effective way and divide the data into groups.
2. Select another subgroup and then redo the first point. Continue the division in each of the following subgroups
3. Continue the breakdown until all observations after the split belong to the same group of target variable or until the stopping condition has been met.

Determining the best division of a subgroup of data depends on the algorithm teaching the decision tree. Among the most popular are CART, C5.0 and CHAID. The first two of them behave in a similar way. Both build full trees, deliberately overfitting the model then cutting off the branches reducing the size of the tree and both are suitable for both categorical and regressive problems. CHAID, in contrast, divides the nodes on the basis of a statistical Chi-square test, so that it can only be applied to categorical variables (Abbott, 2014). In addition, the decision trees may have the following parameters:

- Maximum depth of the tree.
- Minimum number of samples required to split an internal node.
- Minimum number of samples required to be a leaf node. (Pedregosa, 2011)

3.4.3 Artificial Neutral Network

Another algorithm worth mentioning in the supervised learning context is Artificial Neutral network. This algorithm was created to imitate a complicated system of transmitting information that occurs in animal brains. Similarly to natural brains, artificial neutral network consists of

many single neurons. In the case of artificial neural networks, a single neuron behaves like linear regression. An example of a single neuron is shown in figure 19.

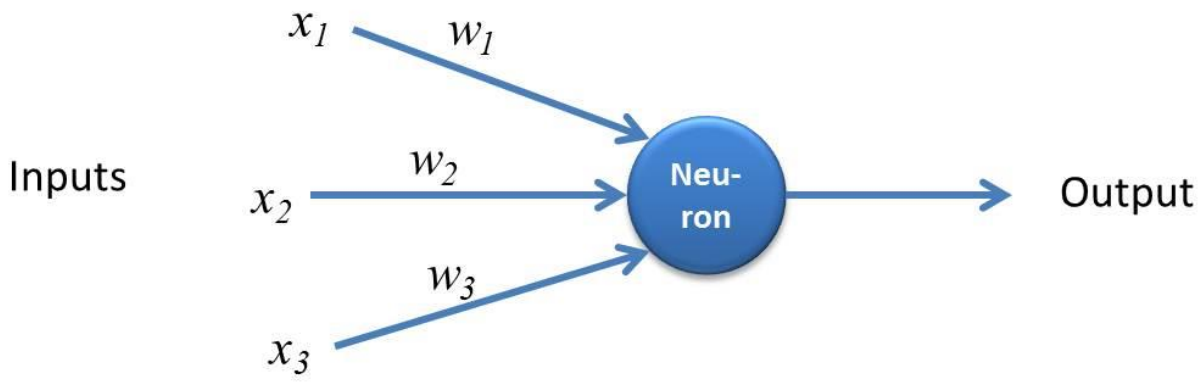

Figure 19 Sigle neuron

Source:https://kindsonthegenius.com/blog/2018/01/what-is-perceptron-how-the-perceptron-works.html

The later output of the linear part of the neuron is transformed using the so-called activation function. It converts the linear result into a binary result, usually using the sigmoid function. The individual neurons are merged to form so-called layers. The layers that provide input data are called input layers and those that form an output layer provide output data. Between them are invisible to users network layers called hidden layers. The parameters are passed from layer to layer until they reach the output layer. Output of neural network can be used to predict both continuous results and binary problems. An example of a neural network is shown in figure 20.

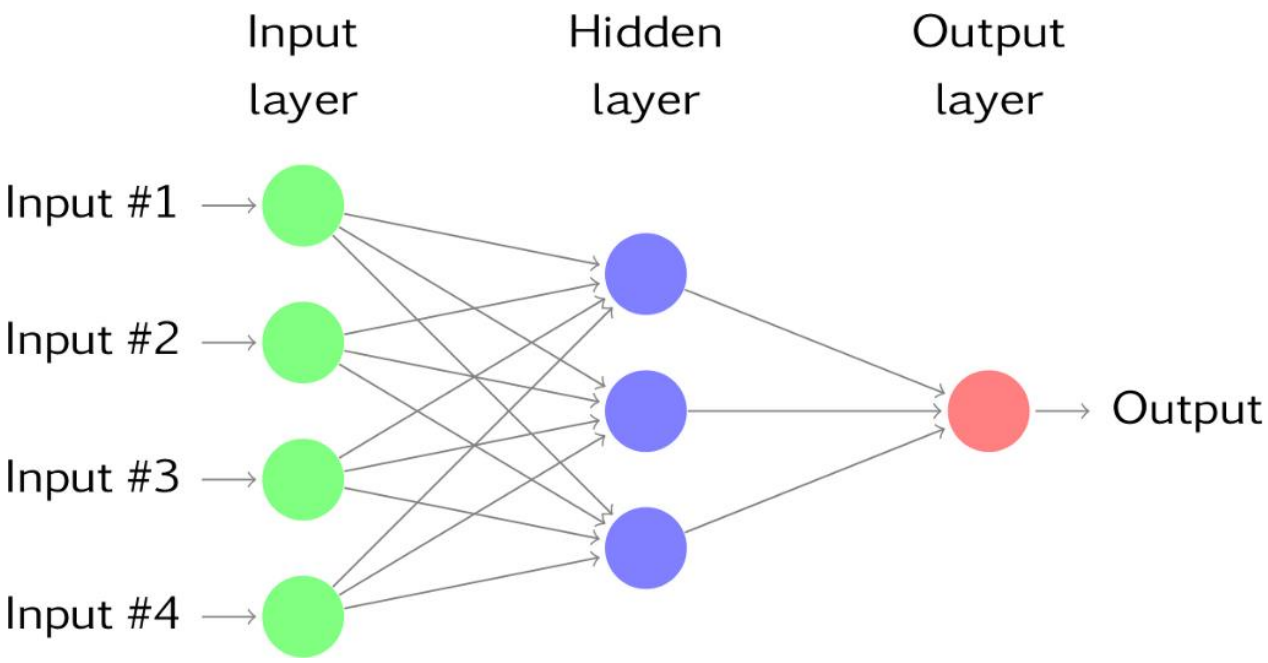

Figure 20 Neutral network

Source: https://mc.ai/artificial-neural-networks-and-deep-learning/

Neural networks are an algorithm for supervised learning, thus using target variable to find the most optimal settings and parameters. During neural network learning, the predicted result of the network is compared with the correct target variable value using such measures of sum of error squared for regression problems and cross-entropy function for classification problems. Optimizing the neural network is therefore about finding network settings that will minimize error functions. Due to the non-linear activation function used in neurons, network optimization is done using the gradient-descent method. This is a method that uses the derivatives of the activation function to indicate the direction in which weights between nodes should be changed. The gradient descent does not provide an exact value, so together with parameters such as the learning rate it creates a process of learning and finding the minimum loss function. The appropriate setting of the learning rate parameter is important in order for the function to be able to find not only the local minimum but also the global minimum loss function. (Larose & Larose, 2015). The iterative learning process of the neural network can be presented in the following steps:

1. Random initialization of network weights.
2. Feed forward the network and calculate loss function.

3. Backpropagate and update the weight of network.
4. Repeat the step 2 until achieving loss function minimum.

Accordingly, single neurons use linear functions, the process of preparing data for artificial neural networks should proceed as with all other numerical algorithms, i.e. all categorical variables must be converted to numerical and missing value must be replaced.

3.4.4 Support Vector Machine

Another supervised learning algorithm used for classification problems is the support vector machine (SVM). It is an algorithm used to solve both regression and classification problems. In case of a classification problem, the aim of SVM will be to develop such a method to effectively separate hyperplane in a multidimensional variable space. Support Vector Machines is in a way an extension of linear classifiers, allowing to find ways to classify data in multidimensional space. The way SVMs extend linear classifiers is to add features such as polynomials of input variables. So adding non-linear features to linear models can make the classifier much more robust. The information about which power to raise the input variable to make the best possible decision boundary is fortunately not a lucky guess, but it can be found with the so-called Kernel trick. It works in such a way that it calculates scalar products for data points for the expected value of the feature without calculating it. An example of decision boundary is presented in figure 21. (Cristianini & Taylor, 2013)

Like other numerical algorithms, Support Vector Machines require the data set to be converted to numerical format and the imputation of missing data. An additional aspect worth mentioning is that SVMs use Eulier distances, so they are very sensitive to data that is not on the same scale. Therefore, Z-scores or Min-Max score should be used before building the model.

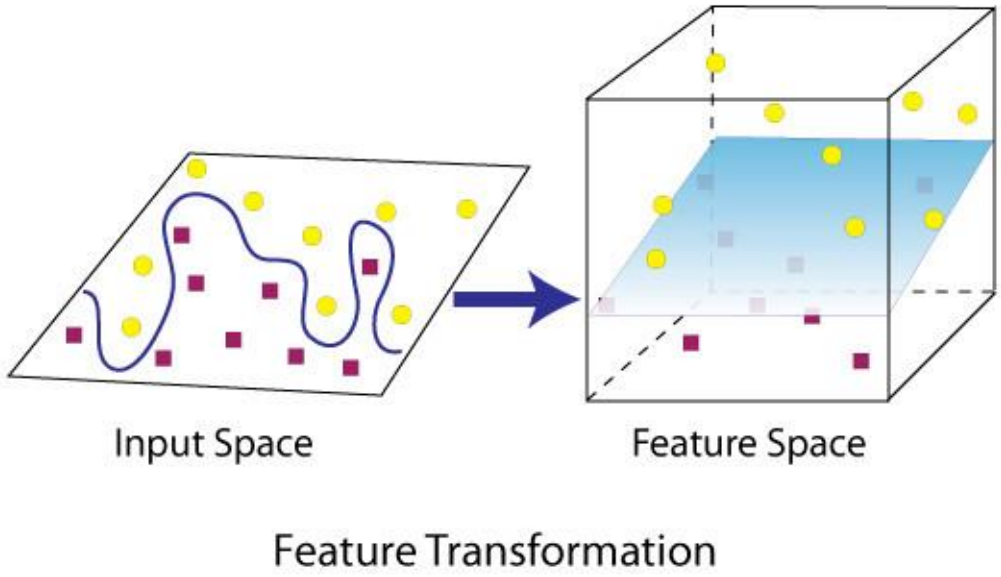

Figure 21 High dimensional hyperplane separation in SVM

Source:https://www.researchgate.net/figure/Demonstration-of-finding-a-separating-hyperplane-in-high-dimensional-space-vs-in-low_fig1_42388347

3.4.5 Methods of improving performance

Having defined supervised learning algorithms, it is also worth mentioning the methods used to improve the effectiveness of the models. The ensembles model means combining two or more models to increase their effectiveness. It has been noted that building an enabled model allows to significantly increase accuracy and is now used in most models in practice. They significantly help to find the optimal choice between bias, i.e. the predictive behavior of the model and variance, i.e. applying it to a new data set.

One method used for this purpose is bagging, which stands for Bootstrap Aggregation. It means building multiple decision trees with bootstrap resembled data and combining predictive results by averaging or voting. The purpose of bagging is overfit models to reduce bias and variance by aggregating multiple decision trees. Despite the fact that this model was originally designed for decision trees, but can also be successfully used in other algorithms.

Boosting is another method used to improve algorithm quality. The principle of its functioning is based on the iterative operation of an algorithm in which misclassified observations increase the cost of errors, after each iteration until they are positively classified. In this way the algorithm is able to adjust its parameters to correctly classify as many observations as possible. After hundreds or thousands of iterations, the final predictions are created on the basis of a weighted average of all models' predictions. Boost is applicable to a large number of algorithms, significantly improving their effectiveness.

The algorithm that uses bagging and boosting in its operation is the Random Forest algorithm. It is a significant implementation of the popular decision tree algorithm. In addition to using the above mentioned techniques, it also uses the random input variable selection method used to build trees, thus forcing trees to search for alternative methods to maximize efficiency. (Abbott, 2014)

3.5 Model's performance evaluation

Having built a predictive model, it is worth considering how to assess its effectiveness. The basic feature of the models' reliability is that they meet their business or research goals. Each domain is characterized by different specifications and requirements, and for this reason results that are unacceptable for one domain may be acceptable for another and vice versa. If our research results are highly dependent on human factors, it can be expected that the model will not be able to cover all examples.

Regression problems can be addressed with the mean square error, the sum of the squares of the differences between the predicted and actual value. In the case of our dataset, the basic evaluation criterion will be accuracy, as a classification problem. It is calculated from the amount of properly qualified data against the total predictions. The Confusion matrix is presented in figure 22. Measures allowing for more in-depth analysis of the results obtained are presented there. Apart from the accuracy, the following measures are possible to calculate:

- Sensitivity – Correctly predicted positive cases
- Specificity – Correctly predicted negative cases
- Precision – Correctness of predicting positive class

- Negative Predictive value – Correctness of predicting negative class
- False alarm – Ratio of negative class classified as positive
- False Dismissal - Ratio of positive class classified as negative

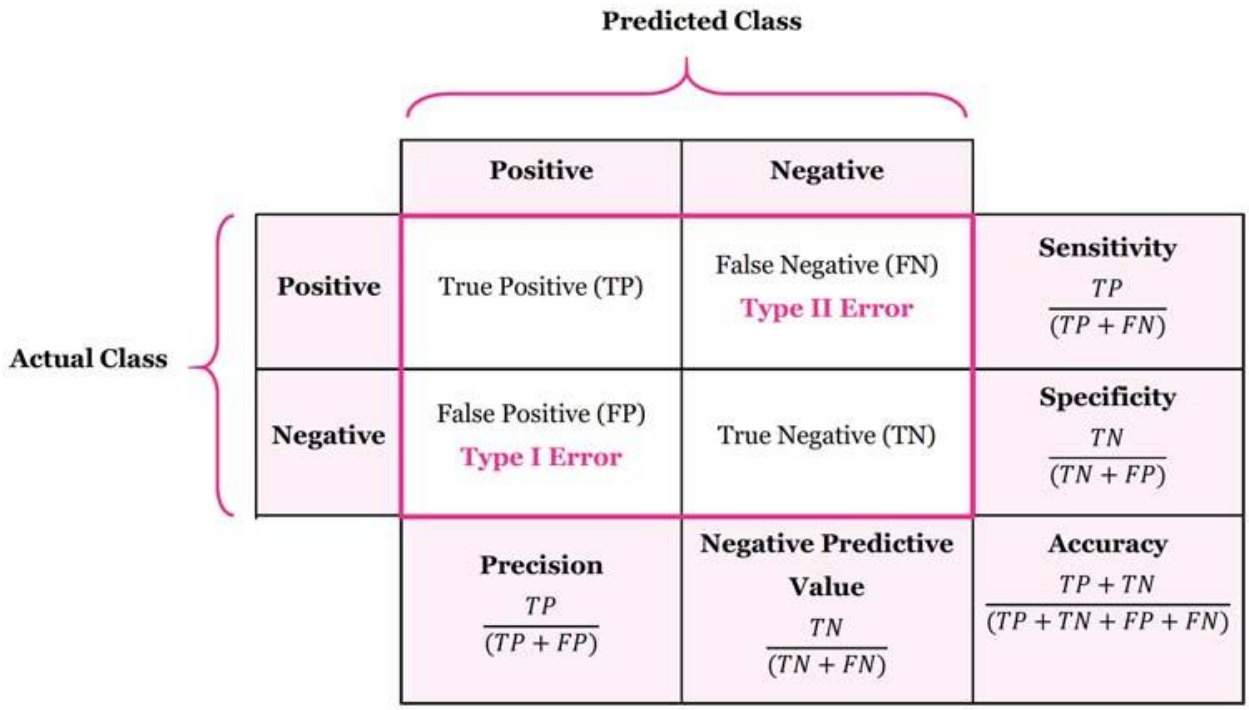

Figure 22 Confusion matrix

Source: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

Receiver Operating Characteristic or so called ROC curve is a measure to visually assess the compromise between false positives ratio on the horizontal axis and true positives ratio on the vertical axis. The comparison of these two values is considered crucial in model assessment. An example of the ROC curve is shown in figure 23. One of the most commonly used measures to evaluate classification models is the measure called Area under the curve, which is a representation of the model performance on the ROC curve. AUC values are in the range [0.1] and higher values are among the more desirable ones. A random classification model will have an AUC of 0.5. The area under the curve is a useful measure as it allows to assess the performance of the model at different stages of the curve. In addition, it is also common practice

to plot several models on one chart, compare their behavior and choose the most advantageous one.

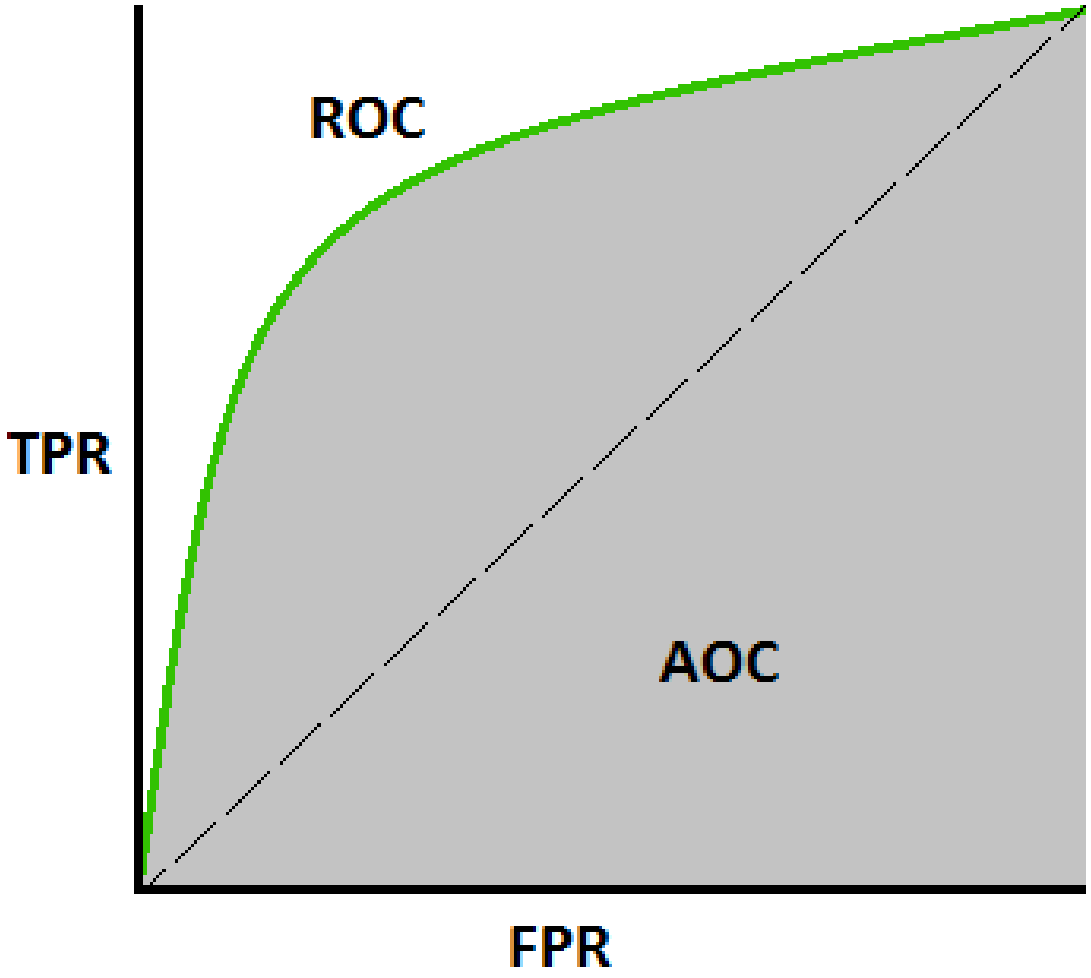

Figure 23 Roc Curve

Source: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

4.Practical Part

4.1 Data set description

The aim of the practical part of the research is to analyze the data set and finally build a model which will help to identify target audience of bank deposit sales. The tools that will be used in

the study are mainly numerical open-source Python libraries such as ski-learn, Pandas and Numpy.

The starting point of the data analysis is to review the data set. Table 3 contains all attributes, with their description and type. For categorical variables, possible categories are provided, whereas numerical type contain unit of measurement.

Variable	**Description**	**Type**
Age	Age	Numeric: Years
Job	Type of job	Categorical
Marital	Marital status	Categorical
Education	Level of education	Categorical
Default	Has credit in default?	Categorical: "no","yes","unknown"
Housing	Has housing loan?	Categorical: "no","yes","unknown"
Loan	Has personal loan?	Categorical: "no","yes","unknown"
Contact	Contact communication type	Categorical: "cellular","telephone"
Month	Contact month of year	Categorical: Months of the year
Day_of_week	Contact day of the week	Categorical: Day of the week
Duration	Last contact duration in sec	Numeric: Seconds
Campaign	Number of contact performed during this campaign	Numeric: Contacts with customer
Pdays	number of days that passed by after the client was last contacted from a previous campaign	Numeric: Days
Previous	number of contacts performed before this campaign	Numeric: Contacts with customer
Poutcome	outcome of the previous	Categorical: "failure","nonexistent","success"

	marketing campaign	
Emp.var.rate	employment variation rate	Numeric: Percentage
Cons.price.idx	consumer price index	Numeric: Index points
Cons.Conf.inx	consumer confidence index	Numeric: Index points
Euribor3m	euribor 3 month rate	Numeric: Percentage
Nr.Employed	number people employeed	Numeric: Number of people
y	has the client subscribed a term deposit?	Binary: "Success" ,"Failure"

Table 3 Original variables

During the construction of the model, however, the number of input variables are going to be reduced as the finally included variables are reflecting the opinions of researchers and analysts described in the review literature section.

According to the authors P. Kotler and G. Armstrong, personal factors that influence consumer decisions are primarily:

- Age and Life Cycle Stage,
- Occupation,
- Economic Situation,
- Lifestyle.

Attributes describing the customer's experience with the product and the company itself are also important factors influencing customer decisions. The following interactions are worth mentioning:

- Perception - The process in which people find, organize information in such a way that it takes on a meaningful perception,

- Learning - means a change in an individual's behavior along with the experience he/she has gained. Learning can take the form of positive or negative experiences that an individual has had in the past with a certain product,
- Beliefs and Attitudes – Describes what a given person thinks about a certain topic, what feelings drive that person, and what is his/her attitude towards given objects.

Macroeconomic factors have a direct impact on the bank deposits. Authors Fisnik Morina & Rufi Osmani confirmed in their research a positive relationship between the interest rate and the level of deposits. Thanks to the use of Fisher's Effect, i.e. the relationship between inflation and interest rate, we are able to assume that rising inflation should have a positive effect on the sale of deposits. Existing relationship between inflation and the unemployment rate, noticed and described as Phillips curve. It states that inflation and unemployment have an inverse relationship. Higher inflation is associated with lower unemployment and vice versa. Taking advantage of this relationship and the knowledge that inflation has a positive impact on the amount of bank deposits, it can be observed that the growing unemployment rate will have a negative impact on the amount of bank deposits. On the basis of the above analyses and the investigated problem characterization, 12 variables were selected from among the original data set. They will then be used as input data for further prediction models. The selected variables are:

1. Age
2. Job
3. Marital
4. Education
5. Housing
6. Loan
7. Campaign
8. Poutcome
9. Emp.var.rate
10. Cons.price.idx
11. Cons.Conf.inx
12. Euribor3m

4.2 Explanatory data analysis

According to the CRISP-DM process, the first phase of the research should bring the understanding of data set. This phase aims to provide as much information as possible about the dataset. Explanatory data analysis is an activity that will help to bring closer the characteristics of features, both categorical and numerical

4.2.1 Categorical data analysis

The first variable to be described is 'y', i.e. the target variable of the current data set. It contains binary information about whether the contact person has decided to buy a bank deposit. In table 4 we can observe the number of successful and unsuccessful contacts. As we can see, the data set is highly not balanced - only 11.3% of contacts with the customer were successful. Additionally, the visual distribution of the y variable is visible on figure 24.

Has the client subscribed term deposit?	**Frequency**	**% of total**
Failure	36548	88.7%
Success	4640	11.3%

Table 4 Target variable

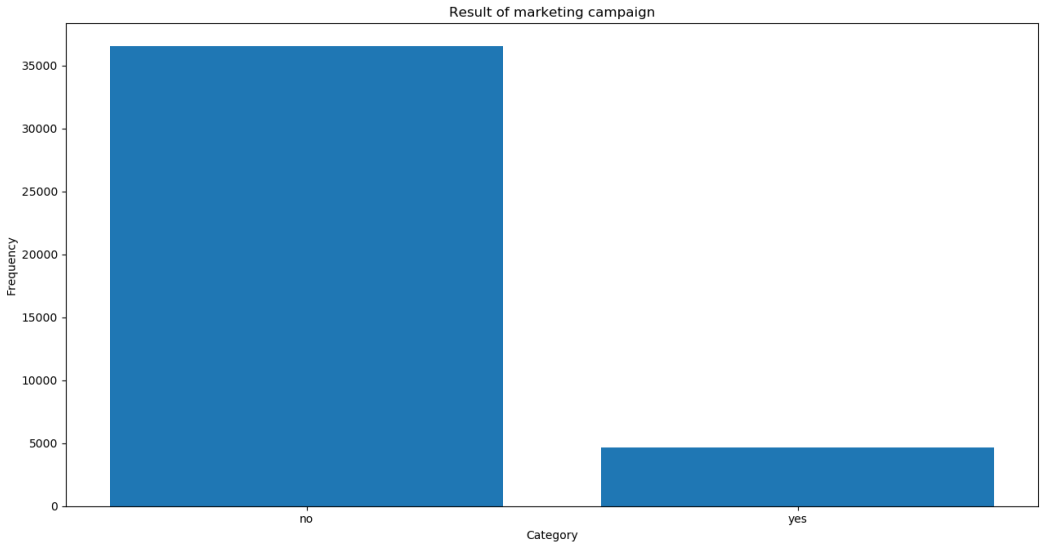

Figure 24 Target variable - bar chart

Source: Author's own work

The next categorical variables described will be input variables. Therefore the distribution of their categories will be shown in the target variable background using contingency tables and stacked bar charts. The first input variable of our data set is 'job'. This variable can distinguish categories of jobs performed by contacted persons. Categories of professions performed together with their counts can be observed in table 5.

Job Category	**Frequency**	**% of total**
admin.	10422	25.3%
blue-collar	9254	22.5%
technician	6743	16.4%
services	3969	9.6%
management	2924	7.1%
retired	1720	4.2%
entrepreneur	1456	3.5%
self-employed	1421	3.5%
housemaid	1060	2.5%
unemployed	1014	2.5%

student	875	2.1%
unknown	330	0.8%

Table 5 Distribution of attribute job

It can be observed that most people contacted during the marketing campaign perform administrative work. Additionally, there are 330 observations whose profession is unknown and have been classified as 'unknown', which constitutes 0.8% of the whole. Table 6 presents contingency table for jobs attribute. Within it we can observe the distribution of campaign results for particular categories of this variable. This allows us to observe which categories turned out to have the best sales performance. Figure 25 shows the bar chart for the job variable. The height of the bar represents frequency counted. You can also see the percentage of the campaign result inside each bar. This allows us to see that in the 'retired' and 'student' categories the deposit sales achieved significantly better results than in others.

Campaign result by job category			
Job Category	Campaign result		
	Success	**Failure**	**Total**
admin.	1355	9067	10422
blue-collar	639	8615	9254
technician	728	6015	6743
services	321	3648	3969
management	327	2597	2924
retired	433	1287	1720
entrepreneur	124	1332	1456
self-employed	149	1272	1421
housemaid	106	954	1060
unemployed	144	870	1014
student	275	600	875
unknown	37	293	330
Total	4639	36549	41188

Table 6 Jobs - contingency table

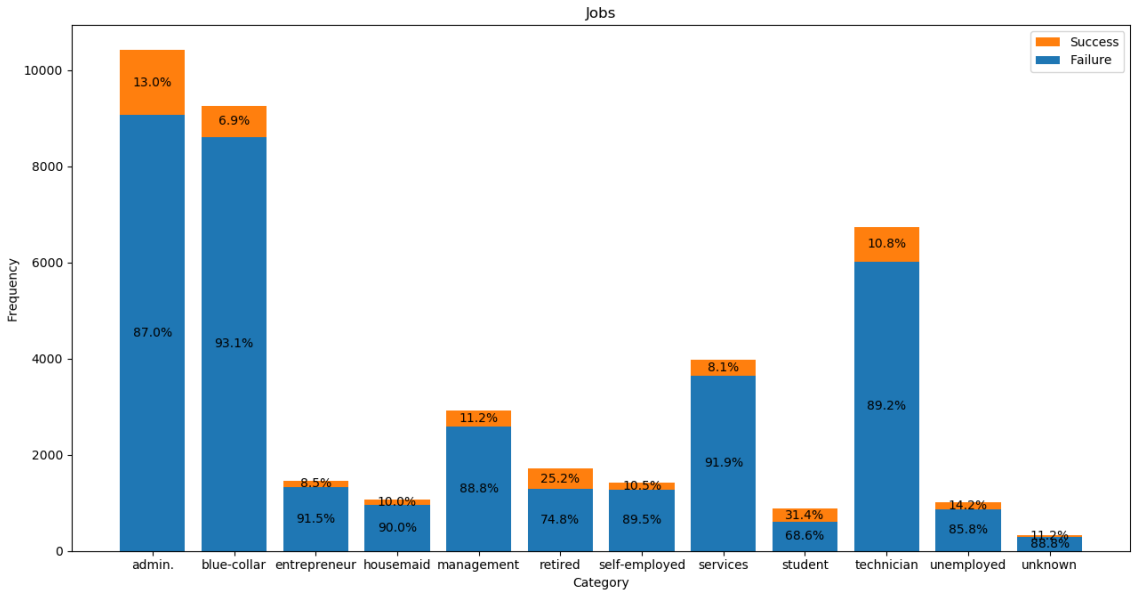

Figure 25 Job - Bar chart

Source: Author's own work

The next categorical variable is 'marital'. It refers to a person's marital status. Table 7 presents the categories of this variable with their frequency. As we can see, the most popular category is 'married', which represents more than 60% of the people being contacted. The status of 80 respondents is unknown, so they have been assigned an unknown category.

Marital	**Frequency**	**% of total**
married	24928	60.5%
single	11568	28.1%
divorced	4612	11.2%
unknown	80	0.2%

Table 7 Distribution of attribute marital

Table 8 presents contingency table for marital status. It presents the distribution of campaign results for each category of the 'marital' variable. Figure 26 shows the bar chart for this variable. The frequency is represented by the height of the bar, while the colors indicate the share of the target variable within the category. As we can see, the 'single' category has a few percent higher share of successful contacts than other categories.

Campaign result by marital category			
Marital Category	Campaign result		
	Success	**Failure**	**Total**
married	2543	22385	24928
single	1620	9948	11568
divorced	475	4137	4612
unknown	12	68	80
Total	4649	36539	41188

Table 8 Marital status - contingency table

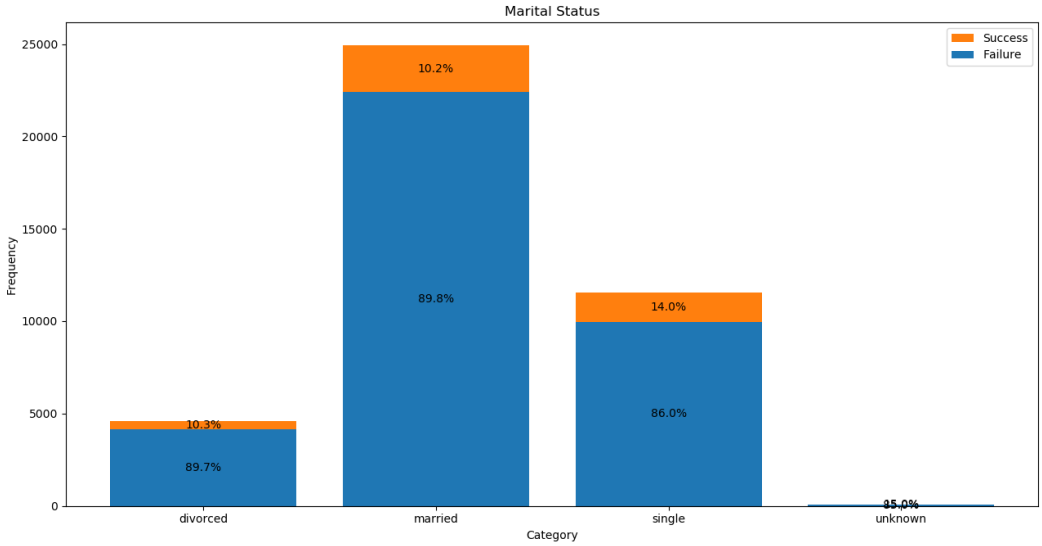

Figure 26 Marital - bar chart

Source: Author's own work

Another analyzed educational variable informs us about the highest level of education achieved by the respondents. Table 9 shows the categories and frequency of this variable. As we can observe, the most popular level of education is the University degree, representing almost 30%. Additionally, the level of education of over 1700 respondents is unknown.

Education	**Frequency**	**% of total**
university.degree	12168	29.5%
high.school	9515	23.1%
basic.9y	6045	14.7%
professional.course	5243	12.7%
basic.4y	4176	10.1%
basic.6y	2292	5.6%
unknown	1731	4.2%

illiterate	18	0.04%

Table 9 Distribution of attribute education

Table 10 shows the contingency table with the results of each category. Additionally, figure 27 shows the frequency of each category along with the proportion of target variable. As we can see, the best sales result was observed for the 'University degree' category.

Campaign result by education category			
Education Category	Campaign result		
	Success	**Failure**	**Total**
university.degree	1667	10501	12168
high.school	1028	8487	9515
basic.9y	472	5573	6045
professional.course	592	4651	5243
basic.4y	426	3750	4176
basic.6y	188	2104	2292
unknown	251	1480	1731
illiterate	4	14	18
Total	4628	36560	41188

Table 10 Education - contingency table

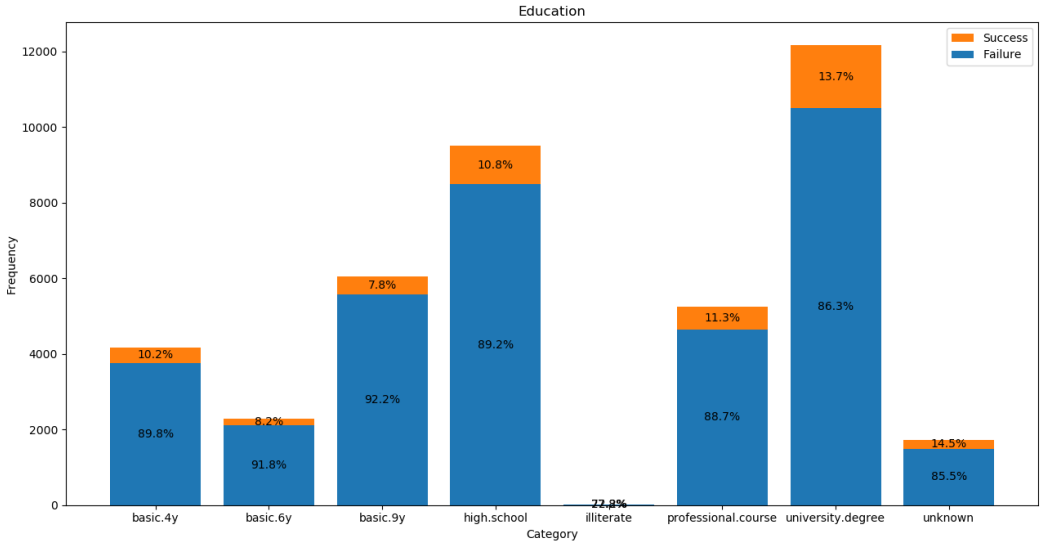

Figure 27 Education - Bar chart

Source: Author's own work

The next variable we will analyze is 'housing'. This attribute provides information on whether the customer in contact has a mortgage. Table 11 shows the categories and number of occurrences of this variable. As we can see, the distribution between clients who have a mortgage is fairly balanced, with a slight advantage of the 'yes' category represented by 52.4% of the total. We can also observe that the value of this variable for nearly one thousand people is unknown.

Has a housing loan?	**Frequency**	**% of total**
yes	21576	52.4%
no	18622	45.2%
unknown	990	2.4%

Table 11 Housing loan

Table 12 contains the categories of the housing loan variable together with the results of the target variable in each category. In addition, figure 28 contains the percentage share of the target variable in the listed categories shown on the bar chart. This allows to observe a slightly higher percentage of successful contacts among mortgage customers - 11.6% vs. 10.9%.

Campaign result by house loan			
House loan	Campaign result		
	Success	**Failure**	**Total**
yes	2503	19073	21576
no	2030	16592	18622
unknown	107	883	990
Total	4640	36548	41188

Table 12 House loan - contingency table

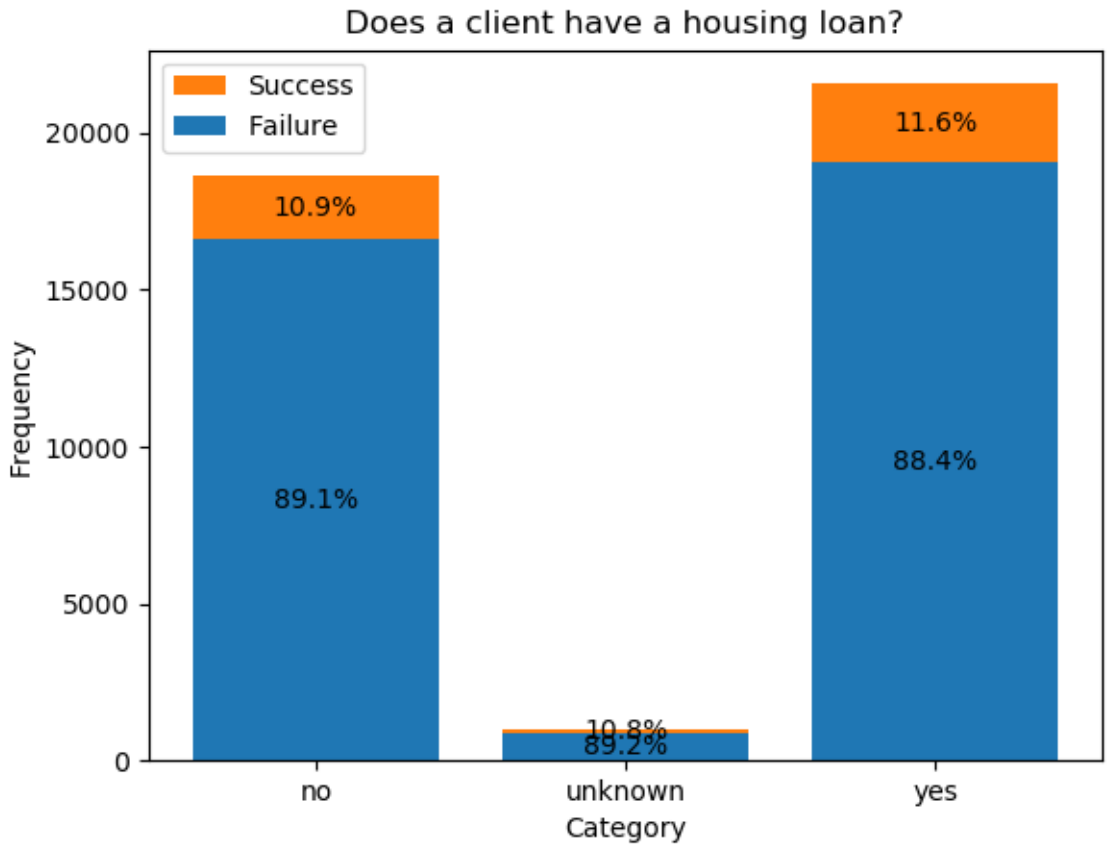

Figure 28 Housing - bar chart

Source: Author's own work

The next categorical variable presented in the data set is 'loan'. This variable indicates whether the contacted person has a personal loan. Table 13 shows the frequency of this variable. As we can see, a significant majority of contacted customers (82.4%) do not have a personal loan. For 990 observations, it was not determined whether the contacted person has a personal loan.

Has a personal loan?	**Frequency**	**% of total**
no	33950	82.4%
yes	6248	15.2%
unknown	990	2.4%

Table 13 personal loan

Table 14 contains a contingency table showing the distribution of campaign results for different categories, while figure 28 presents a bar chart with their percentage share. This allows us to

observe a small advantage of successful contacts among customers without personal credit - 11.3% vs. 10.9%.

Campaign result by personal loan			
Personal loan	Campaign result		
	Success	**Failure**	**Total**
No	3836	30114	33950
Yes	681	5567	6248
Unknown	107	883	990
Total	4624	36564	41188

Table 14 Personal loan - contingency table

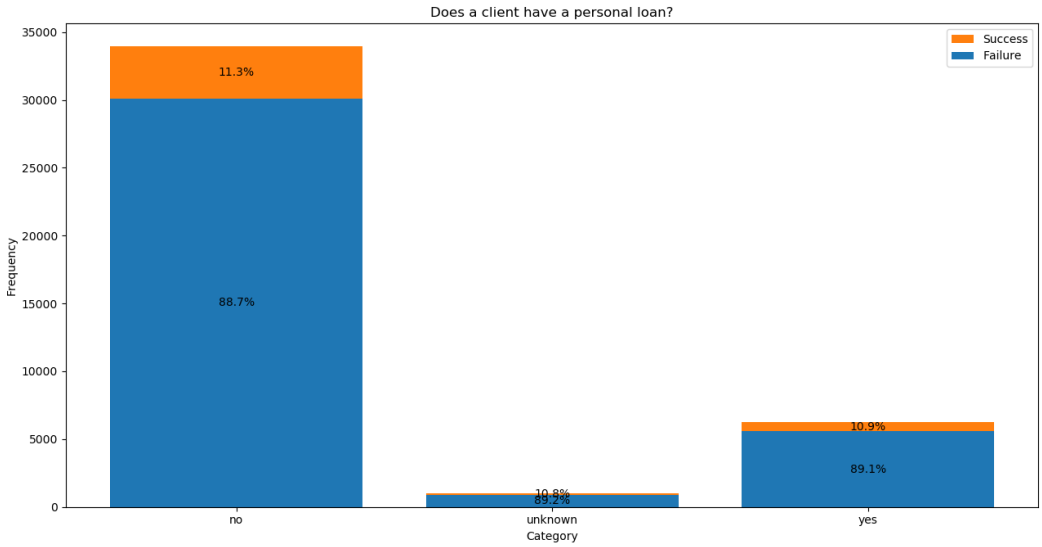

Figure 29 Personal loan - bar chart

Source: Author's own work

The next categorical variable from the data set under consideration is 'poutcome'. It represents the result of the previous marketing campaign of the respondent. The variable assumes the following three categories: 'failure', 'nonexistent', 'success' As we can observe in table 15 the overwhelming majority of customers contacted in the current campaign have not been contacted before and it stands for 86.3%. 3.3% of the customers contacted during the current campaign also purchased the product during a previous campaign.

Previous campaign outcome	Frequency	% of total
Nonexistent	35563	86.3%
Failure	4252	10.3%
Success	1373	3.3%

Table 15 Previous campaign outcome

Table 16 shows the distribution of results of the current campaign according to the results of the previous campaign. As we can see, customers who decided to buy the product in the previous campaign responded remarkably well in the current campaign. This is the only category in which frequency for 'success' exceeded frequency for 'failure'. Additionally, on figure 30 we can observe the percentage share of target variable. Thanks to this, we can see that customers who were contacted during the previous campaign are more willing to buy products than those who were not. This is confirmed for both positive and negative results of the previous campaign.

Campaign result by previous campaign result			
Previous campaign	Campaign result		
	Success	**Failure**	**Total**
Nonexistent	3130	32433	35563
Failure	604	3648	4252
Success	894	479	1373
Total	4627	36561	41188

Table 16 Poutcome - contingency table

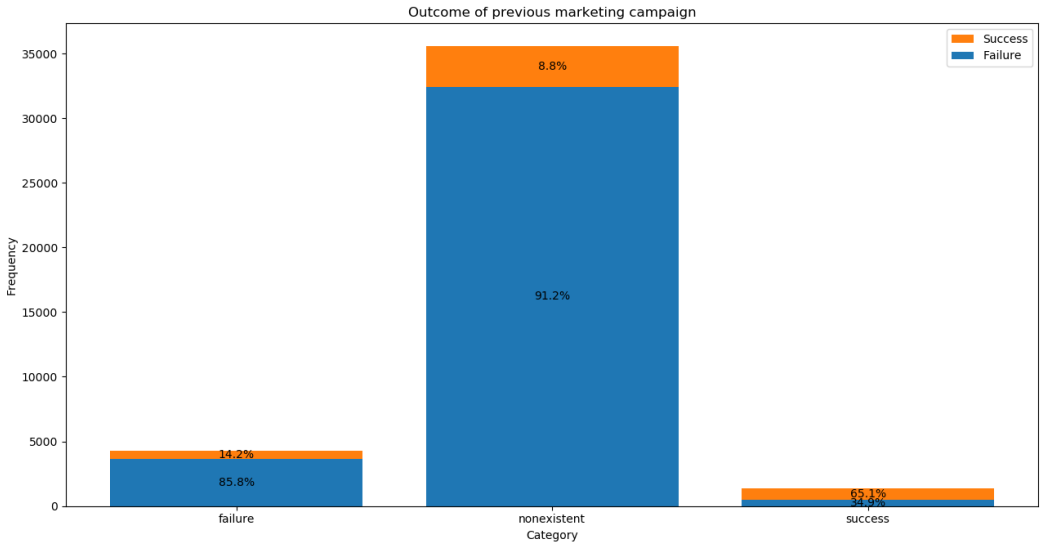

Figure 30 Poutcome - bar chart

Source: Author's own work

4.2.2 Numerical variable analysis

The next step that will be taken during the explanatory data analysis is to examine the numerical variables. Currently, in our selected dataset we have the following numerical variables: age, campaign, emp.var.rate, cons.price.idx, cons.conf.inx and euribor3m. Each of these variables will have descriptive statistics calculated, then its distribution will be checked on the histogram, and then the distribution will be compared for different target variable categories using a box plot. Finally, the correlation of all variables will be checked using r-Pearson coefficient. This will allow us to detect the flaws in the distribution, make a first look at what numbers we are dealing with and compare how the values of individual variables are presented for different values of the target variable. Correlation analysis of variables will allow us to determine whether there is a 'multicollinearity' phenomenon present in the data set.

The first numerical variable to be analyzed is 'age'. It indicates the age of the contact person. Table 17 shows descriptive statistics of this variable informing about the distribution parameters. On the basis of this information, we can see that the average age of the contacted person was 40 years, with a standard deviation of about 10 years. On the basis of observations of distribution parameters: mode, mean and median, we can observe that the distribution is probably right-skewed This is also confirmed by the value of the skewness parameter and the shape of the distribution presented in figure 31. The value of exceed kurtosis is more than 0, which means that we can observe platykurtic in the given distribution. Figure 31 also shows a comparison to a normal distribution. As can be seen, apart from a positive skew, the distribution takes a shape similar to normal distribution.

Parameter	**Value**
Mode	31
Mean	40.02406
Standard deviation	10.42125
Min	17.00

25%	32.00
50%	38.00
75%	47.00
Max	98.00
Skewness	0.784696
Kurtosis	0.791311

Table 17 Age variable descriptive statistics

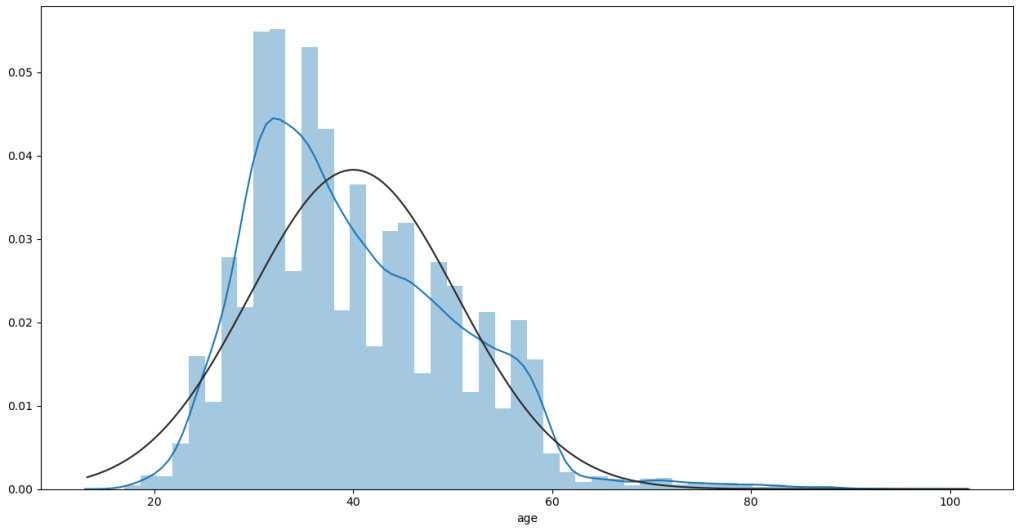

Figure 31 Age – histogram

Source: Author's own work

Figure 32 shows a comparison of the distribution of the age variable for different 'y' results. As we can see the box plot for the value 'yes' takes higher box body boundary. This may mean that on average people of a higher age were more likely to buy deposits.

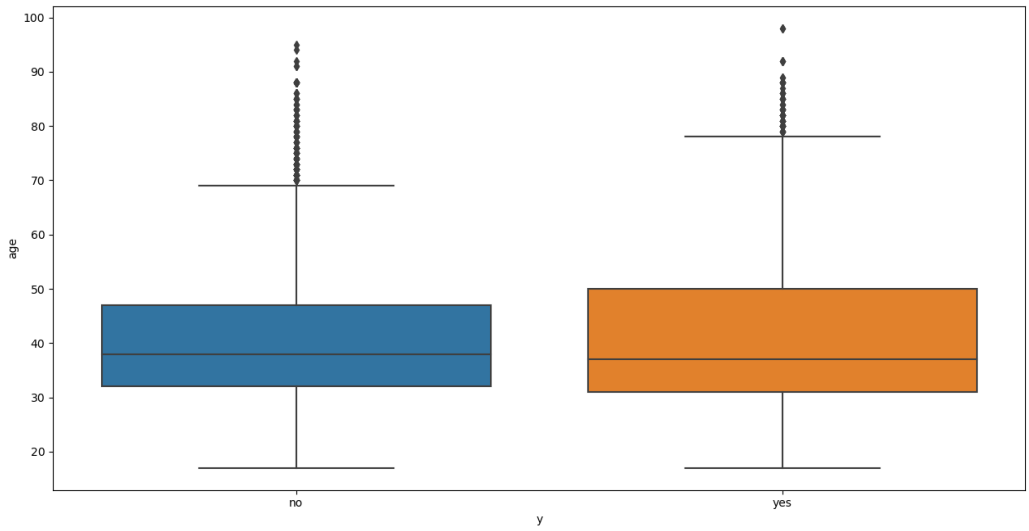

Figure 32 Age - Box Plot

Source: Author's own work

The next variable we will analyses is 'campaign'. This means how many times a person has been contacted during current campaign. Table 18 shows descriptive statistics of this variable. As we can see, the person usually was contacted once, but there were also cases of more frequent contacts. The maximum number of contacts for one person was 56 times. As we can observe from the skewness parameter the distribution is clearly right-skewed. The high parameter of the kurtosis also means that the distribution of the variable is noticeably concentrated in its tails. Figure 33 shows a histogram on which we can see the distribution of the campaign variable. As was shown in the mode, the peak of the distribution occurs at a value of 1, and the extreme values make the distribution clearly right-skewed. Figure 34 shows the distribution of the number of contacts for different values of the target variable. As we can see, the body box for successful contacts is slightly lower, and the upper pin pointing 1.5 IQR is much lower than for failed ones. From this information we can conclude, that more contacts are not beneficial for a higher probability of success.

Parameter	Value
Mode	1
Mean	2.567593
Standard deviation	2.770014
Min	1.00
25%	1.00
50%	2.00
75%	3.00
Max	56.00
Skewness	4.76251
Kurtosis	36.9798

Table 18 Campaign variable descriptive statistics

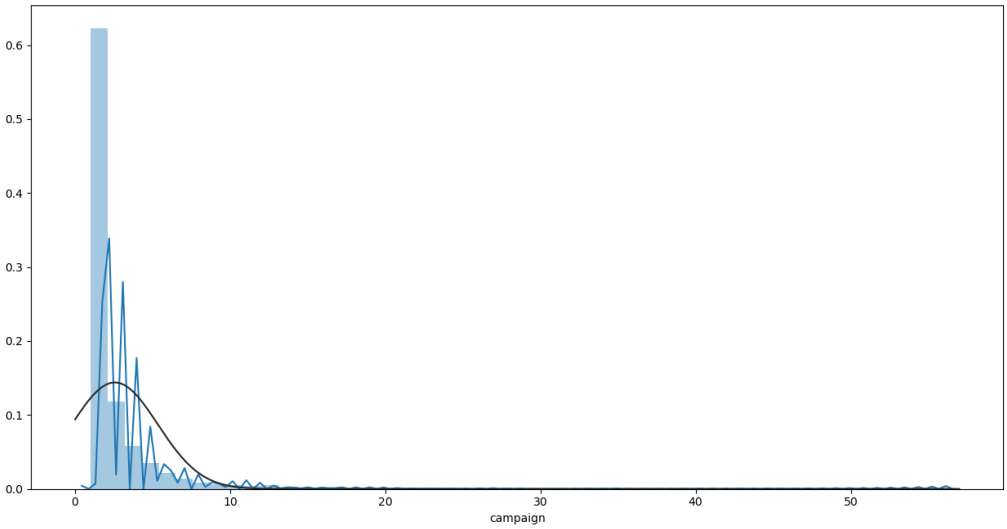

Figure 33 Campaign – histogram

Source: Author's own work

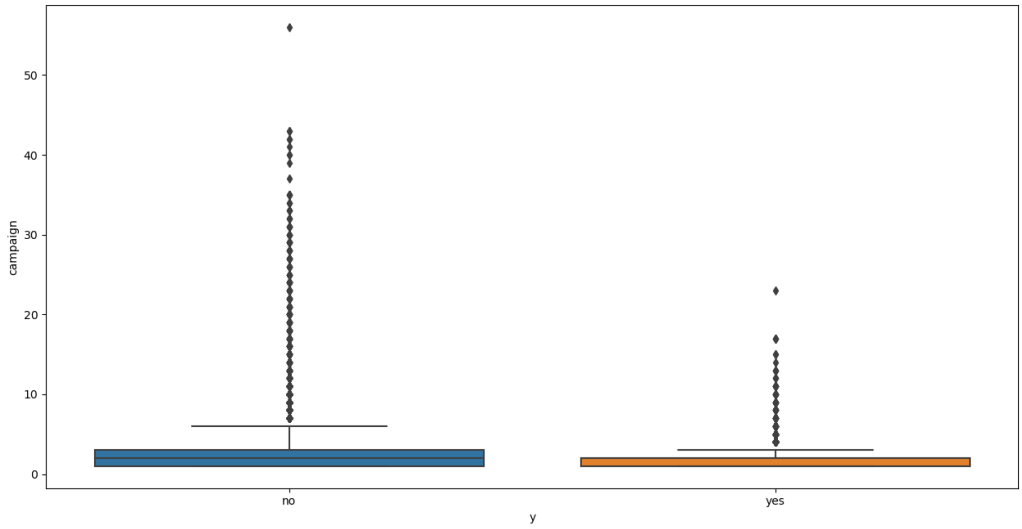

Figure 34 Campaign - box plot

Source: Author's own work

The next variable discussed is the macroeconomic variable emp.var.rate. This means the quarterly employment variation rate in Portugal. Table 19 shows descriptive statistics for this variable. As we can see, it takes values from -3.4 to 1.4. Negative values mean a decrease in employment, while positive values mean an increase in employment.

Parameter	**Value**
Mode	1.4
Mean	0.081886
Standard deviation	1.570960
Min	-3.40
25%	-1.80
50%	1.10
75%	1.40
Max	1.40
Skewness	-0.724096
Kurtosis	-1.062632

Table 19 Emp.var.rate variable

From figure 35 we can see that this variable should probably not be treated as numerical, because its values take only a few levels, which makes us see several peaks in the histogram, without any continuity between them. This indicates that it is suitable for transformation at a later stage of data cleaning. The box plot visible on figure 36 indicates a significant difference in the employment rate values against the two target variable values. It is noticeable that more often than not, cases ending in successive sales occurred for negative employment variation rate values.

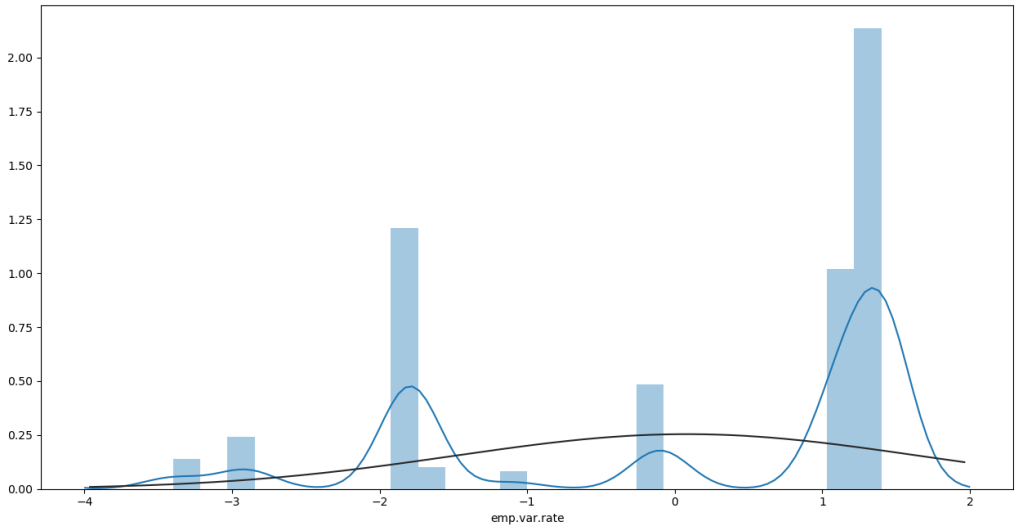

Figure 35 Emp.var.rate – Histogram

Source: Author's own work

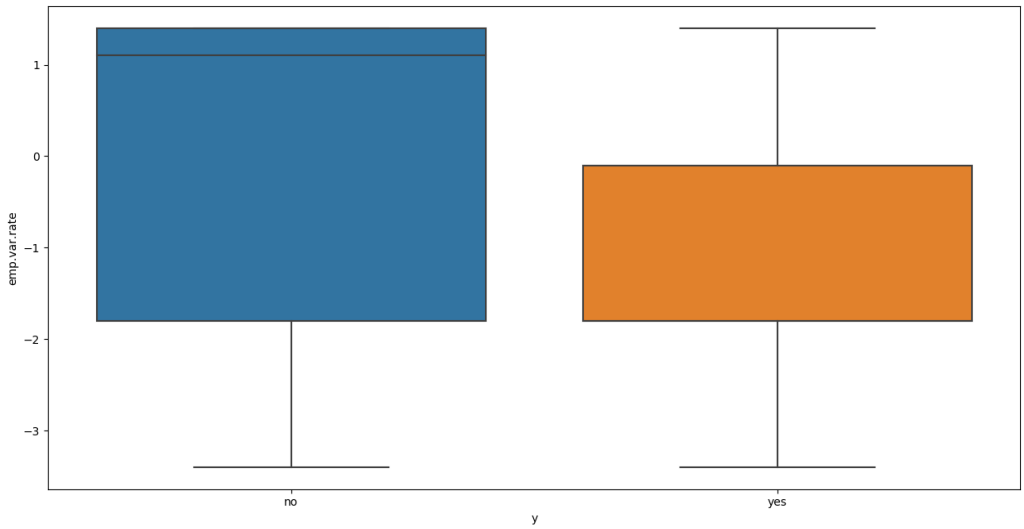

Figure 36 Emp.var.rate - Box plot

Source: Author's own work

The following numerical variable under analysis is cons.price.idx. It represents the monthly value of the Consumer Price Index, which informs us about the inflation rate. The base year used to calculate the CPI in this case is 2012, so as we see in table 20 all values are below 100. Figure 37 shows a histogram showing the distribution of this variable. As we see, it presents a multi-modal distribution with a few major spikes. Probably this variable will be an object of categorization during data preparation. Figure 38 shows a box plot for the CPI variable. As we can see, bank deposit sales went a little better for lower CPI values.

Parameter	**Value**
Mode	93.994
Mean	93.576
Standard deviation	0.5788
Min	92.20
25%	93.08
50%	93.75
75%	93.99
Max	94.77

Skewness	-0.2309
Kurtosis	-0.8298

Table 20 Cons.price.idx variable

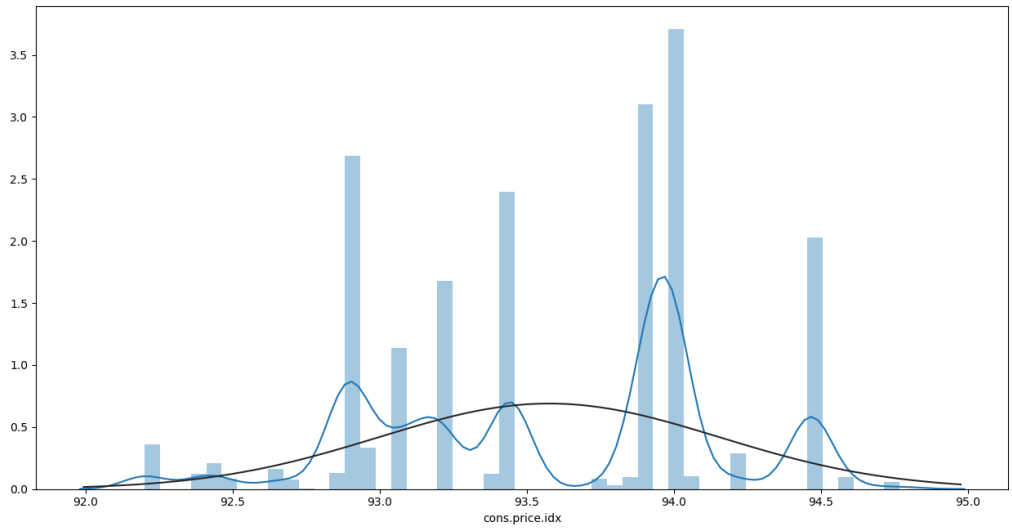

Figure 37 Cons Price idx – Histogram

Source: Author's own work

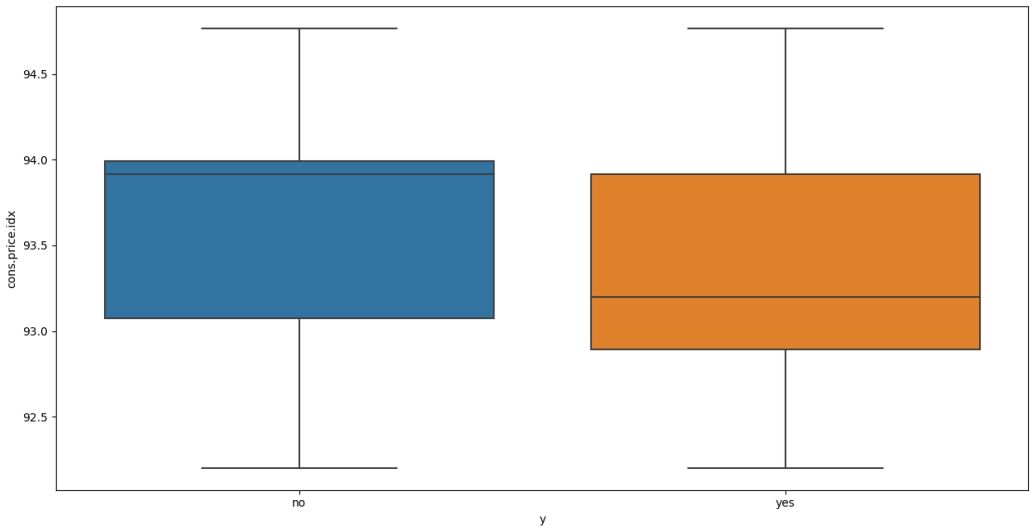

Figure 38 Cons Price idx - Box plot

Source: Author's own work

The next variable analyzed is cons.conf.idx. It represents the monthly value of the consumer confidence index in Portugal. The value of this index is calculated on the basis of interviews with customers about their opinions about the economic future of the national economy as well as their purchasing tendencies. The index is calculated using the difference between positive and negative responses. As we can see in Table 21, the value of the CCI for each period during the marketing campaign takes a negative value, which may result from the global financial crisis occurring at this time. Just as we can see on figure 40, the CCI distribution for successful cases is different from that for failed ones. We can say that as consumer morale fell, consumers were more likely to buy a bank deposit.

Parameter	**Value**
Mode	-36.4
Mean	-40.50
Standard deviation	4.6281
Min	-50.80
25%	-42.70
50%	-41.80
75%	-36.40
Max	-26.90
Skewness	0.3032
Kurtosis	-0.3586

Table 21 Cons Conf Inx variable

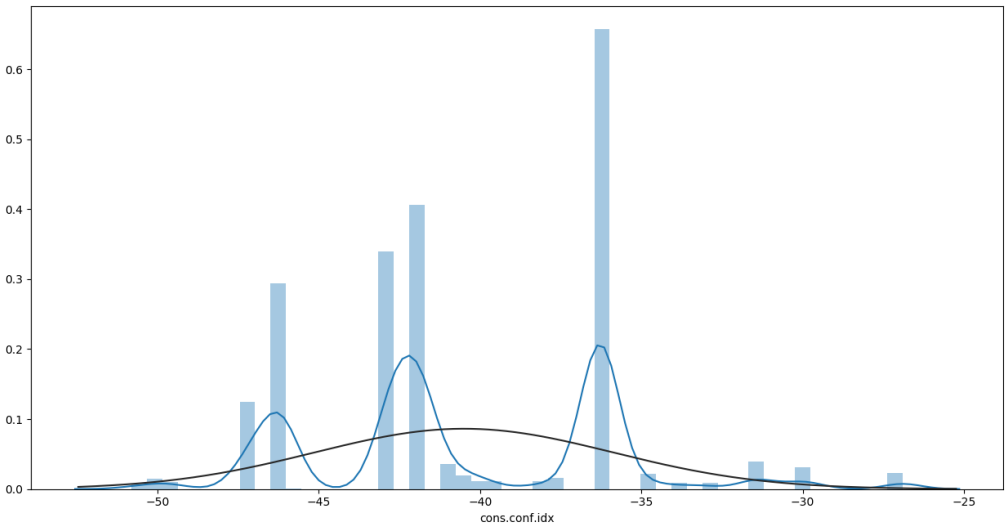

Figure 39 Cons Conf Idx – Histogram

Source: Author's own work

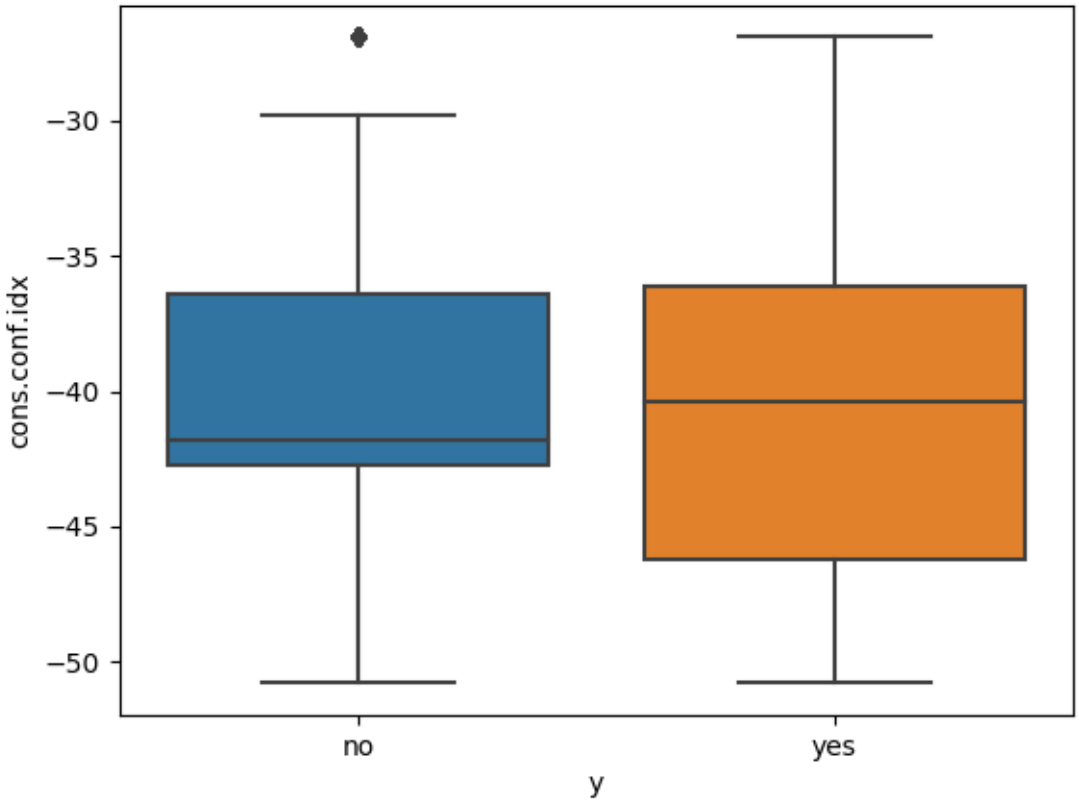

Figure 40 Cons Conf idx - Box plot

Source: Author's own work

The last numerical variable to be subjected to explanatory analysis is 'euribor3m'. This indicator stands for the Euro Interbank Offered Rate, which is the rate at which banks lend to each other short-term. This rate has a very significant impact on the value of interest rates on deposits offered to individual customers. The skewness parameter indicates a left-skewed distribution. As we can see in figure 41, the variable distribution takes a bi-modal shape. The first peak appears in the area of -2 and the second 1. This variable is likely to be subject to transformation during the data cleaning process. Figure 42 helps to observe differences in the distribution of the variable against the target variable. As we can see, the lower values of Euribor3m translated into better deposit sales during the campaign.

Parameter	**Value**
Mode	4.857
Mean	3.6213
Standard deviation	1.7344
Min	0.63
25%	1.34
50%	4.86
75%	4.96
Max	5.05
Skewness	-0.7092
Kurtosis	-1.4068

Table 22 Euribor3m variable

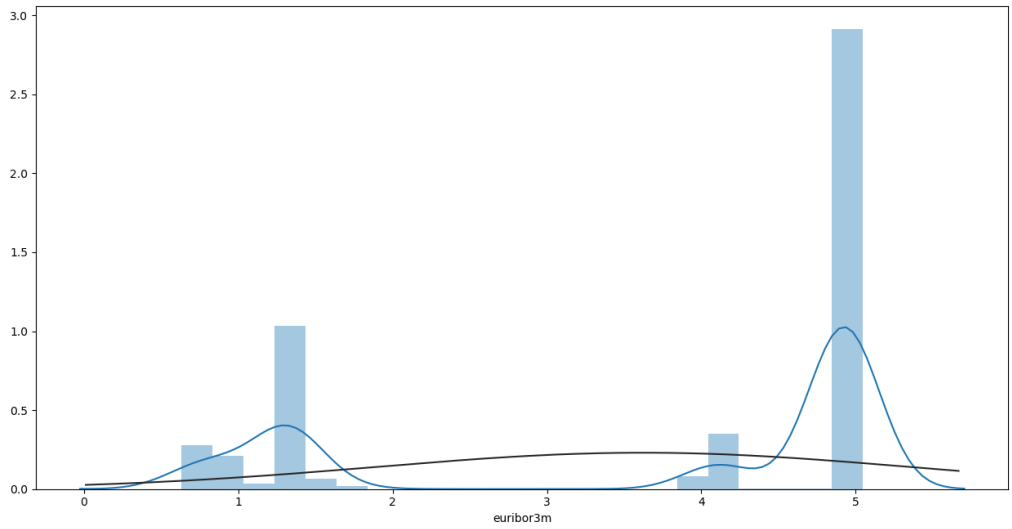

Figure 41 Euribor3m – histogram

Source: Author's own work

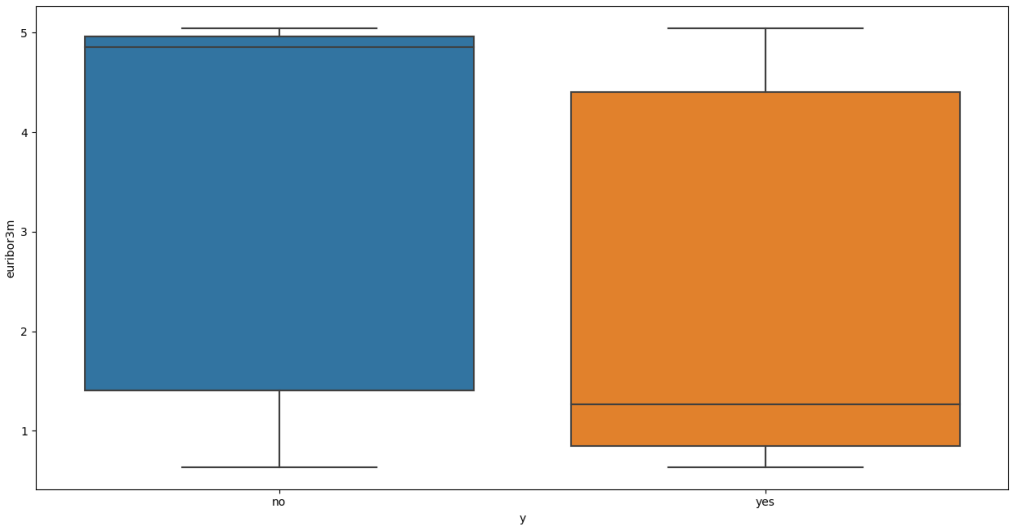

Figure 42 Euribor3m - box plot

Source: Author's own work

After we have analyzed the numerical variables and evaluated their descriptive statistics as well as the distribution of univariate and target variable, in order to detect the linear dependencies of the input variables, we will analyses their correlation. The phenomenon 'multicollinearity'

means a strong linear relationship between the model input variables. A strong correlation is considered to be a correlation parameter value above 0.8. If the input variables contain such a strong correlation, appropriate steps should be taken because the information contained in these variables will be duplicated. Figure 43 contains the correlation matrix for numerical variables. As we can see, the correlation coefficient between the variables emp.var.rate and euribor3m is 0.97. This is a very strong coefficient and indicates that both variables move practically collinearly. This means that one of these variables can be removed because it does not bring new information. In addition, a strong correlation is noticeable between emp.var.rate and cons.price.idx of 0.78. On the basis of these coefficients, the emp.var.rate variable will be excluded from the data set because the information contained in this variable is contained in other variables.

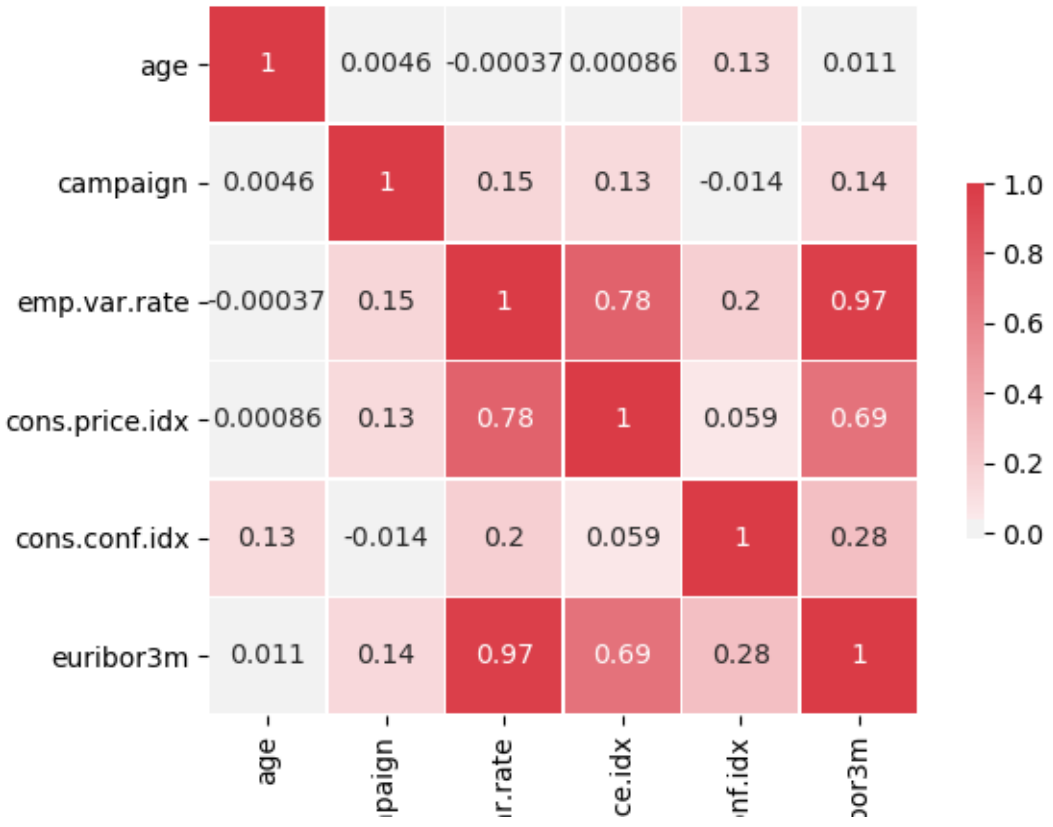

Figure 43 Input variable correlation

Source: Author's own work

4.3 Data preparation

Studies on data understanding have identified problems that will be corrected at a subsequent phase to provide the possibility of data modelling. Following the CRISP-DM methodology, the next step after Data Understanding is Data Preparation. During the previous phase, flaws in the data were detected, which we will now try to correct. These are:

- Unknown category existing in some attributes,
- Multicollinearity amongst input variables. To prevent this phenomena variable ‘ emp var rate’ is going to be excluded from the data set,
- Skewness of variables age will be corrected by logarithm transformation.

In addition to correcting existing attributes, new features will be created from some variables. Due to bi-modal or multi-modal distributions of some variables, they will be binned, i.e. categorical variables will be created from them. These will be:

- Poutcome,
- campaign,
- emp.var.rate,
- cons.price.idx,
- cons.conf.inx,
- euribor3m

4.3.1 Missing data handling

In the case of variables used in our dataset, missing data represent a marginal percentage. They occur in the case of several categorical variables and are represented as 'unknown' category. As Table 23 indicates, the share of missing data in a particular variable does not exceed 5%, and for some variables takes exactly the same value, which may suggest that missing information in one variable may also not be present in others. So in order to remove the missing value we will use listwise deletion of each observation that contains at least one missing value.

Variable	% of missing
Job	0.8%
Marital	0.2%
Education	4.2%
Housing	2.4%
Loan	2.4%

Table 23 [%] of missing data

As a result of list wise deletion, observations containing missing value have been removed from the data set. As a result, the filtered data set contains 38 245 observations. This means that the size of the new dataset is 92.85% of the size of the original dataset, and the filtered observations represented 7.15%.

$$^{38245}/_{41188} = 92.85\%$$

4.3.2 Fixing linear dependencies

To avoid the phenomenon of 'multicollinearity' and its consequences, highly linearly correlated variables should be removed from the data set because they contain duplicate values. As discovered in figure 42, there is a really strong correlation between several variables. The strongest relationship visible between emp.var.rate and euribor3m is 0.97, which means a nearly perfect linear relationship of these variables. The next strong correlation is emp.var.rate and cons.price.idx, with a correlation coefficient of 0.78. For this reason, the emp.var.rate variable will be excluded from later analysis and the whole data set.

4.3.3 Skewness correction

As we saw during the explanatory data analysis part, most of the numerical variables had very clear peaks and their distribution was not continuous, so they will be binned into categorical variables. However, the age variable has a continuous distribution which, apart from a strong

right-sided obliqueness, has a distribution close to normal. For this purpose, the values of this variable have been transformed ln(1+x), thus creating the age_log variable. We can see the distribution of the age_log variable on figure 44, it is much closer to normal than the original variable (visible on figure 31).

$$Skewness\ age = 0.8040$$

$$Skewness\ age\_log = 0.1739$$

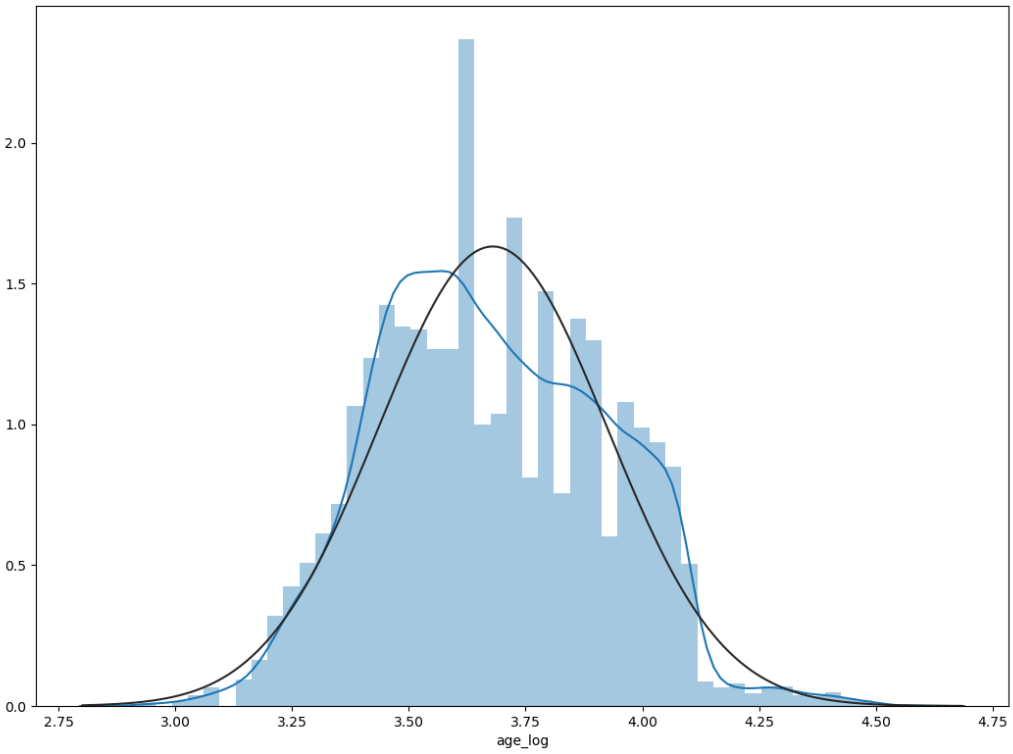

Figure 44 Age_log distribution

Source: Author's own work

4.4 Modelling

According to the CRISP-DM process, the post-fixing data phase is model building. However, to do this it is necessary to create the input data in a form that is suitable for a specific model. Algorithms that will be used during modelling are logistic regression and decision tree. They have been selected for their analysis and interpretation capabilities, but each of them requires input in a different form. Therefore, from now on, changes made to the data set will be made separately for both models.

4.4.1 Preparing data set – Logistic regression

Logistic regression is an algorithm that can only process numerical variables. For this purpose, all categorical variables of our data set must be encoded as numbers. For variables such as 'education', this is not a problem because it is an ordinal variable, so successive levels of education can be assigned to consecutive digits. However, the other variables do not have a fixed order, so they cannot be treated in this way. Therefore, n-1 dummy variable will be created from the following n categories. In this way the following changes have been made for logistical regression purposes:

- Contact_cat – It was created from the transformation of the variable 'campaign'. The variable takes the value of 1 when there was one contact, 2 when there were two contacts and 3 when there were more than two contacts..
- From the 'poutcome' variable, n-1 binary variables are created
- A binary variable 'eurlibor3m >3' has been created from 'eurlibor3m'.
- Cons.price.idx was divided into categories starting with a value of 1 for CPI <93 and increasing by 1 for each subsequent value of 0.5 CPI up to 94.5. Values above 94.5 were coded as '5'.
- Cons.conf.idx has been similarly divided into categories. A value of 1 means CCI <-45 and increases by 1 for each subsequent 5 CCI value to -35. Values above -35 have been coded as '4'.
- The variables 'house' and 'loan' were encoded as binary variables.

- From the variable 'marital', n-1 binary variables have been created
- An ordinal variable was created from the 'education' variable. The first three levels have been merged into one category, so the higher level of the variable means the higher level of education achieved.

The original variables were then replaced by new categorical variables. The list of variables and their coding can be found in appendix 1.

4.4.2 Building a model – Logistic regression

After the data set was prepared for logistic regression. We can start building the model. First, the original data set was divided into training and testing parts in the proportion of 3:1. Training part of data set will be used to calculate the model parameters, while testing part will be used to verify the accuracy of the model. This is important to verify the model for data that was not used for modelling.

Then, a model was built on the training set. To balance the target variable classes the parameter class_weight='balanced' of the Logistic Regression model scikit-learn was used. Logistic regression parameters stimulated by maximum likelihood method were presented in table 24.

Variable	**Parameter**	**P-value**
Const	-0.3536	0.1809
Age_log	-0.1001	0.1325
poutcome_nonexistent	0.4038	0.0000
poutcome_success	1.9118	0.0000
eurlibor3m_above_3	-2.0317	0.0000
Contact_cat	-0.0589	0.0001
CPI_cat	0.2095	0.0000
CCI_cat	0.3260	0.0000
personal_loan	-0.1026	0.0049
mortgage	-0.0102	0.7000
marital_married	0.0288	0.5052

marital_single	0.1484	0.0028
job_blue-collar	-0.1527	0.0008
job_entrepreneur	0.0531	0.4690
job_housemaid	-0.2046	0.0268
job_management	0.0844	0.1283
job_retired	0.5117	0.0000
job_self-employed	0.0345	0.6370
job_services	-0.2118	0.0000
job_student	0.3642	0.0014
job_technician	-0.0487	0.2352
job_unemployed	0.0028	0.9751
Edu_cat	0.0690	0.0000

Table 24 Logistic regression full data set parameters

As we can see, some of the parameters in Table 24 have a p-value above 0.05, which for the Student's significance t test means that the parameter is not significant. Parameters that are not significant should also be excluded from the model. Thus, the logistic regression model has been built once again, containing only relevant parameters. Table 25 shows the parameters of the rebuilt model.

Variable	**Parameter**	**P-value**	**Impact**
Const	-0.7057	0.0000	Negative
poutcome_nonexistent	0.4053	0.0000	Positive
poutcome_success	1.9080	0.0000	Positive
eurlibor3m_above_3	-2.0306	0.0000	Negative
Contact_cat	-0.0600	0.0001	Negative
CPI_cat	0.2095	0.0000	Positive
CCI_cat	0.3227	0.0000	Positive
personal_loan	-0.1047	0.0040	Negative
marital_single	0.1361	0.0000	Positive

job_blue-collar	-0.1474	0.0003	Negative
job_housemaid	-0.2194	0.0147	Negative
job_retired	0.4672	0.0000	Positive
job_services	-0.2061	0.0000	Negative
job_student	0.3961	0.0004	Positive
Edu_cat	0.0725	0.0000	Positive

Table 25 Logistic regression significant parameters

Through the estimation of parameters and subsequent verification of their significance as well as rebuilding of the model, we were able to find not only variables that have a significant impact on whether the contact person will buy a bank deposit, but also the value of individual parameters. Due to the parameters obtained during the estimation, we are also able to calculate the odds ratio, visible on equation 1 and the probability of the event, visible on equation 2. It should be pointed out that event 1 is a deposit sale success.

$$\frac{P(1)}{1-P(1)} = -0.7057 + 0.4053\ \llbracket poutcome \rrbracket\_nonexistent + 1.9080\ \llbracket poutcome \rrbracket\_success - 2.0306\ eurlibor3m\_(above\_3) - 0.0600\ Contact\_cat + 0.2095\ CPI\_cat + 0.3227\ CCI\_cat - 0.1047\ personal\_loan + 0.1361\ marital\_single - 0.1474 job\_bluecollar - 0.2194\ job\_housemaid - 0.4672\ job\_retired - 0.2061\ job\_services + 0.3961\ job\_student + 0.0725\ Edu\_cat$$

Equation 1 Odds ratio

$$P(1) = 1/(1 + e\ \hat{}\ -(-0.7057 + 0.4053\ \llbracket poutcome \rrbracket\_nonexistent + 1.9080\ \llbracket poutcome \rrbracket\_success - 2.0306\ eurlibor3m\_(above\_3) - 0.0600\ Contact\_cat + 0.2095\ CPI\_cat + 0.3227\ CCI\_cat - 0.1047\ personal\_loan + 0.1361\ marital\_single - 0.1474\ job\_bluecollar - 0.2194\ job\_housemaid - 0.4672\ job\_retired - 0.2061\ job\_services + 0.3961\ job\_student + 0.0725\ Edu\_cat)$$

Equation 2 Probability of event

By analyzing the impact of given variables on the result, it is possible to see which variables increase the chances of buying a bank deposit. Thus, looking at Table 25, we see that among the personal characteristics, the level of education and the information that a person is single increase the chances of success. Among the professions, job_student and job_retired increase

the likelihood of sales, while job_blue-collar, job_housemaid and job_services have a negative impact on it.

In terms of macroeconomic factors, we can see that sales went better when the consumer price index and consumer confidence index were higher and Eurlibor3m was below 3%.

Among consumer factors, we can see that personal loan had a negative impact on deposit sales, and the lower number of contacts supported better sales. In addition, the fact that the customer has been contacted in the past positively influences the result of current sales.

4.4.3 Preparing data set – Decision Tree

Decision trees are algorithms that deal well with both numerical and categorical data. For this reason, decision tree models do not require restrictive changes to the data set to create a model. In addition, numerical variables do not need to be assumed to have a normal or uniform distribution. Thus, during modelling, original numerical variables can be used without the need to transform them for correcting skewness or outliers influence.

Nevertheless, decision trees through their structure performing recursive splitting activities are somehow exposed to too many breakdowns when n levels of categorical variables are transformed into n-1 binary variables as was the case with logistical regression, the so-called one-hot encoding. This leads to blocking of individual parts of the tree, limiting its propagation to a particular branch and a general deterioration in accuracy. An example of a one-way tree is shown in figure 45.

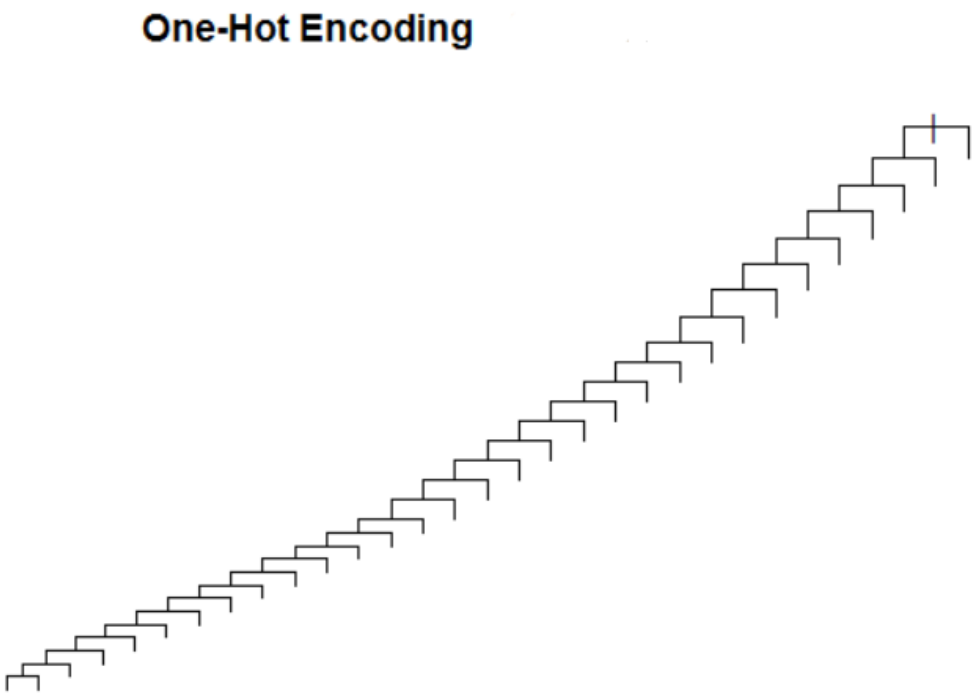

Figure 45 One-Hot encoding decision tree

Source: https://medium.com/data-design/visiting-categorical-features-and-encoding-in-decision-trees-53400fa65931

For this reason, categorical variables will not be encoded using the One-Hot encoder method, but numeric encoding will be used, i.e. a number has been assigned to each category, allowing the tree to grow to similarity to that of figure 46.

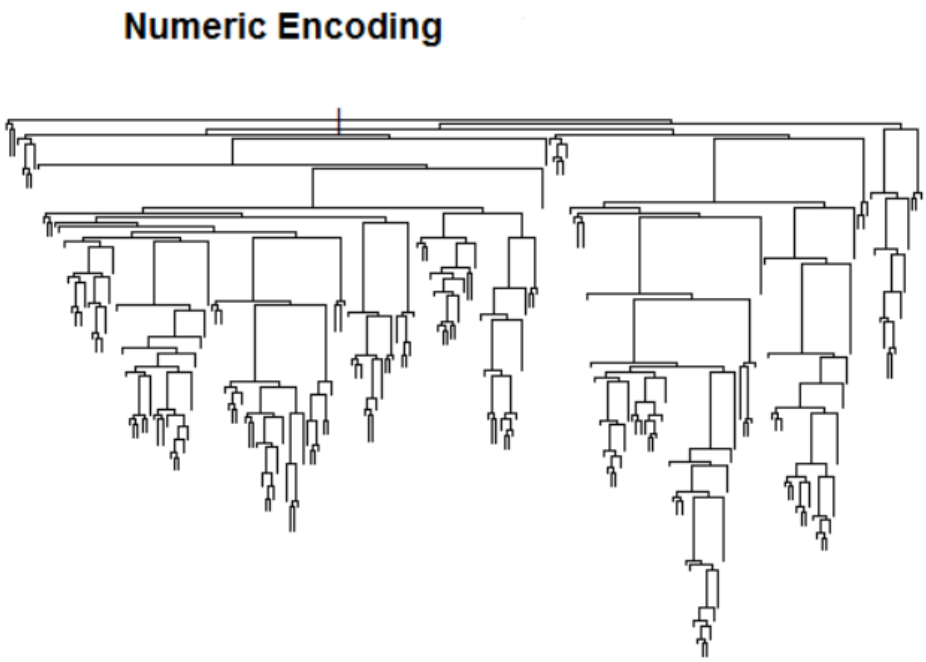

Figure 46 Numeric Encoding - decision tree

Source: https://medium.com/data-design/visiting-categorical-features-and-encoding-in-decision-trees-53400fa65931

The changes that have been made to the data set for the purpose of building the decision tree model are as follows:

- Job_Cat – Numeric encoding for 11 levels of variable ‘job’
- Contact_cat – It was created from the transformation of the variable 'campaign'. The variable takes the value of 1 when there was one contact, 2 when there were two contacts and 3 when there were more than two contacts.
- CPI_cat - Cons.price.idx was divided into categories starting with a value of 1 for CPI <93 and increasing by 1 for each subsequent value of 0.5 CPI up to 94.5. Values above 94.5 were coded as '5'.
- CCI_cat - Cons.conf.idx has been similarly divided into categories. A value of 1 means CCI <-45 and increases by 1 for each subsequent 5 CCI value to -35. Values above -35 have been coded as '4'.

- Marital_cat –Numeric encoding for 3 levels of variable ‘marital’
- Edu_cat – Numeric encoding for 4 levels of variable ‘education. Original categories representing different levels of primary school has been binned into one category.
- Outcome_cat – Numeric encoding for 3 levels of variable ‘poutcome’. 1 stands for non-existing, 2 for failure and 3 for success.
- Rate_cat - This variable was created by binning the original 'euribor3m'. 6 categories have been created, with 1 being the value of 'euribor3m' below 1, category 2 being a rate between 1 and 2 and so on. Category 6 represents a rate above 5.
- The variables 'house' and 'loan' were encoded as binary variables.

The original variables were then replaced by new categorical variables. The list of variables and their coding are enclosed in appendix 2.

4.4.4 Building a model – Decision Tree

After building a data set appropriate for the decision trees, we can start to actually build the model. First, we have split the data set into training and test groups. The training data set will be used to build the model, whereas the predictive behavior of the model will be tested on the test group. This is important because a reliable assessment of the model must be made on data that the model has not seen before. First, the decision tree has been built using all the features of the data set. The result is the tree visible in figure 47.

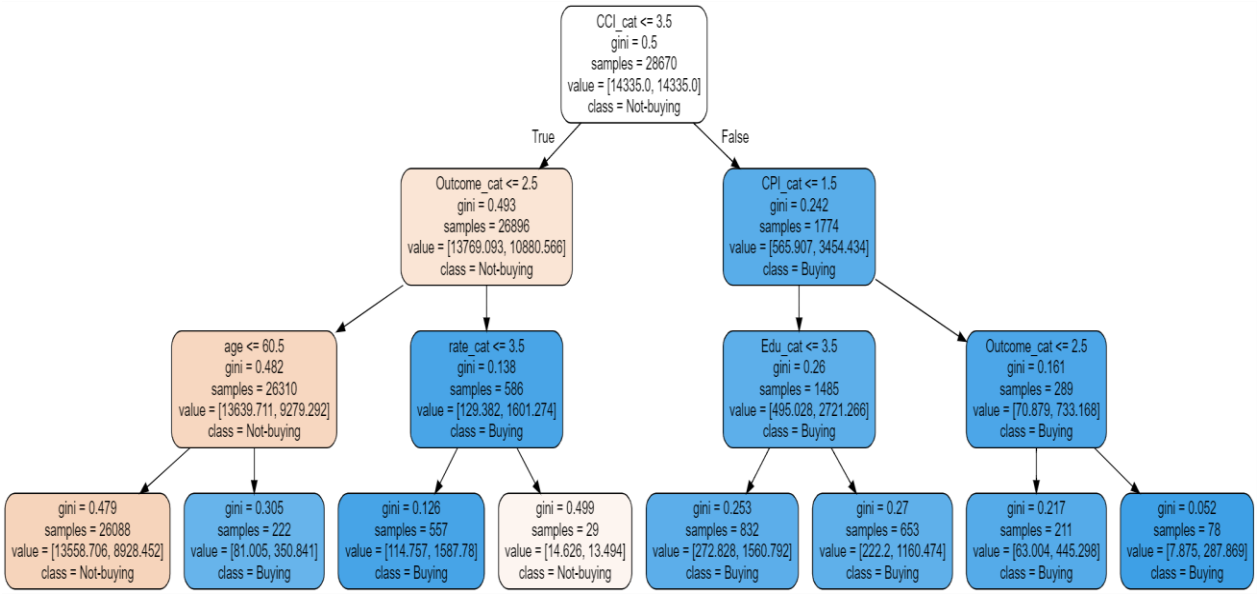

Figure 47 Decision Tree - All features

Source: Author's own work

As we can observe, a formed tree has a sequence of conditions that lead to the classification of the observation into the appropriate category. The nodes representing the success of a deposit sale are marked in blue as shown in the picture. There is also detailed information about the gini index used for dividing the nodes and the number of samples in each node. To the left of the splitting nodes is the condition 'True' and to the right is 'False'. The arrows intuitively indicate the nodes meeting the condition. As we can see, the depth of the tree is three levels, including root node. This, as well as some other parameters concerning the minimum amount in leaf node or the minimum amount in parent nodes were specified during the construction of the tree to avoid overfitting. Additionally, due to the need to balance the unequal target variable, the cost of misclassification of a minority category has been increased.

Root split is the most important and influential place to split a tree. On the basis of this variable the later divisions of the data set are shaped. In the case of our tree, if the observation has a CCI_cat value above 3.5, it will go into the basket on the right side, and if not, on the left side. Subsequent conditions will then be checked in the next nodes terminal, until the observation goes to leaf node. In the case of the right side from root split all leaf nodes indicate the same class (buying), which is not a rule. Therefore, we'll analyze the conditions to the left of the root split. Therefore, if the CCI_cat variable was below 3.5 then the outcome of previous campaign

is checked. If the previous campaign ended with an outcome other than success (Outcome_cat <= 2.5), the value of the variable 'age' will be checked. If the client's age is below 60.5 then it goes to 'Non-buying', otherwise 'buying. However, it should be remembered that the conditions from the previous nodes terminal must still be maintained. Table 26 shows the combinations of conditions from each leaf nodes and their results. The numbering of leaf nodes counts from left to right.

Leaf Node	**Conditions**	**Result**	**No. of samples**
1	CCI_cat <= 3.5 Outcome_cat <= 2.5 Age <= 60.5	Non-buying	26088
2	CCI_cat <= 3.5 Outcome_cat <= 2.5 Age > 60.5	Buying	222
3	CCI_cat <= 3.5 Outcome_cat > 2.5 Rate_cat <= 3.5	Buying	557
4	CCI_cat <= 3.5 Outcome_cat > 2.5 Rate_cat > 3.5	Non-buying	29
5	CCI_cat > 3.5 CPI_cat <= 1.5 Edu_cat <= 3.5	Buying	832
6	CCI_cat > 3.5 CPI_cat <= 1.5 Edu_cat > 3.5	Buying	653
7	CCI_cat > 3.5 CPI_cat > 1.5 Outcome_cat <= 2.5	Buying	211
8	CCI_cat > 3.5	Buying	78

	CPI_cat > 1.5 Outcome_cat > 2.5		

Table 26 Tree 1 splitting conditions

Although the decision tree was built from all the available variables, as was seen in figure 46, not all features were equally important during splitting nodes. Features' importance analysis is aimed at finding the most important variables that are most commonly used for splitting nodes. As we can see in figure 48, there are only a few variables in our decision tree that are really important in the modelling context. The two most frequently used are CCI_cat and Outcome_cat, which prove that the macroeconomic environment and consumer history are important issues, and the bank's sales department should bear this in mind. In addition, a much lower share, but still above the others, has a variable age. It seems that bank deposits sell better in particular age groups. In the context of the model, this leads us to conclude that these are the most important variables, mostly responsible for the outcome of the research.

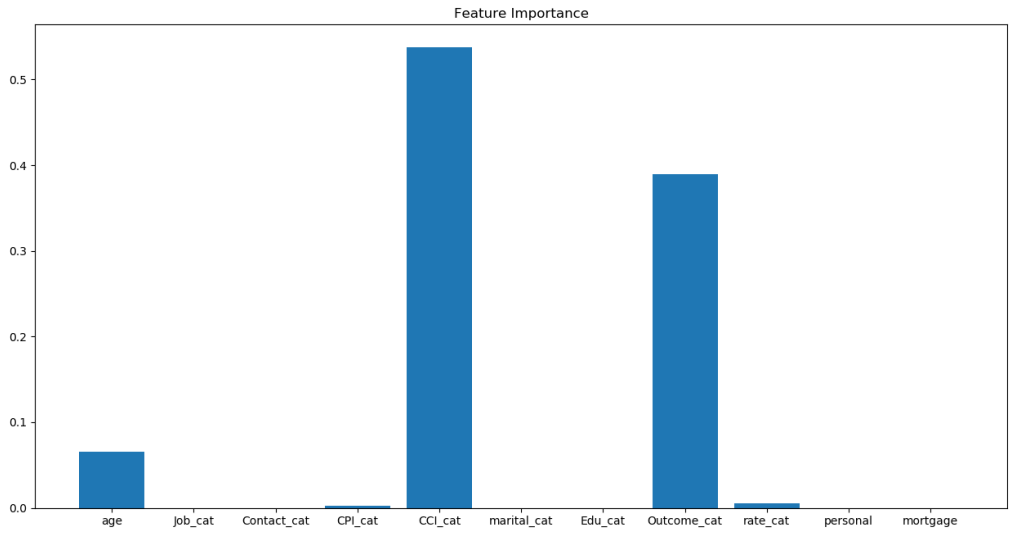

Figure 48 Tree 1 features importance

Source: Author's own work

Features' importance analysis gives a picture of which variables are really important for prediction. This allows to rebuild our model. This means building it from the variables that have

a significant impact on the target variable. So, the decision tree visible in figure 49 consists only of age, CCI_cat and Outcome_cat, which are the most important variables.

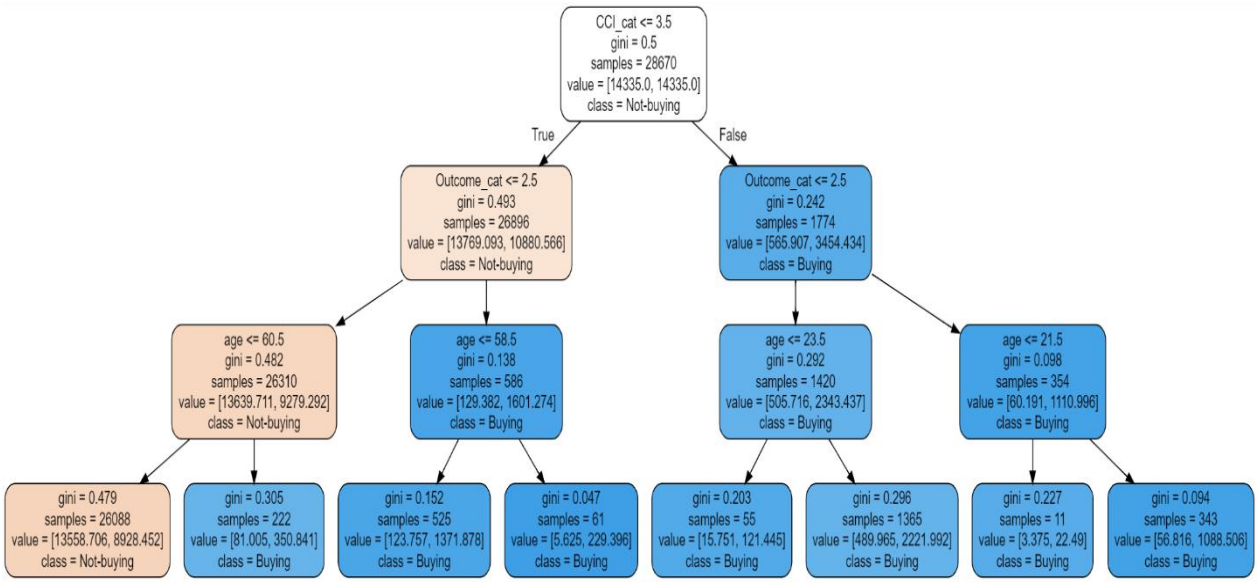

Figure 49 Decision Tree – Best features

Source: Author's own work

As we can observe the structure and shape of the tree, it looks very similar to tree 1. This means that in fact most of the model variants were explained by these three variables. Splitting rules for tree 2 are enclosed in table 27.

Leaf Node	**Conditions**	**Result**	**No. of samples**
1	CCI_cat <= 3.5 Outcome_cat <= 2.5 Age <= 60.5	Non-buying	26088
2	CCI_cat <= 3.5 Outcome_cat <= 2.5 Age > 60.5	Buying	222
3	CCI_cat <= 3.5 Outcome_cat > 2.5 Age <= 58.5	Buying	525

4	CCI_cat <= 3.5 Outcome_cat > 2.5 Age > 58.5	Buying	61
5	CCI_cat > 3.5 Outcome_cat <= 2.5 age <= 23.5	Buying	55
6	CCI_cat > 3.5 Outcome_cat <= 2.5 age > 23.5	Buying	1365
7	CCI_cat > 3.5 Outcome_cat > 2.5 age <= 21.5	Buying	11
8	CCI_cat > 3.5 Outcome_cat > 2.5 age > 21.5	Buying	343

Table 27 Tree 2 Splitting Conditions

As in the case of tree 1, we can observe that CCI_cat and Outcome_cat variables are most often used to segregate observations into individual nodes. We can observe that during splitting nodes, the algorithm tried to search for patterns inside the data, the combination of which gave results in successful deposit sales. Some of these connections, visible in table 26 and 27 illustrate what we guessed during explanatory data analysis, when we looked into which categories (for categorical variables) or distribution (for numerical variables) have a higher percentage of successful sales. As we can see, some of these leaf nodes represent values corresponding to students or retired people, those who have responded positively to previous marketing campaigns, and those who were contacted when CCI index was higher.

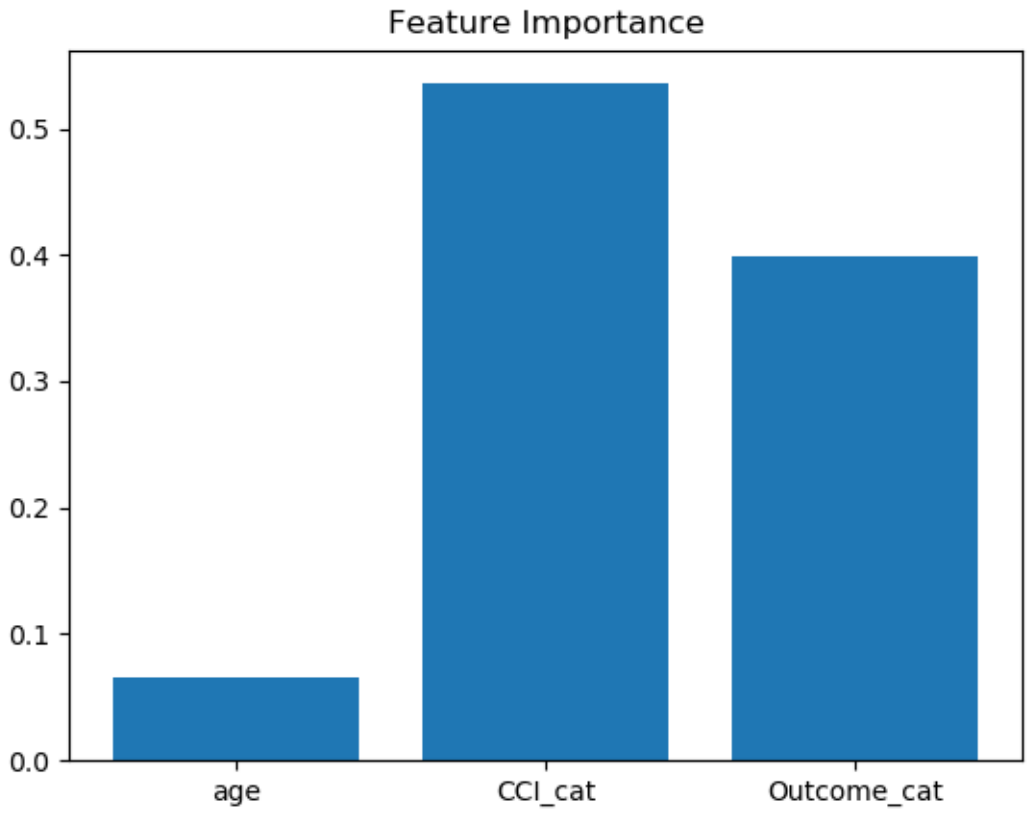

Figure 50 Tree 2 feature importance

Source: Author's own work

Although the CCI_cat variables, Outcome_cat works very well as predictors of client performance, by their predominant importance they limit the possibility of using other variables to categorize observations. And yet these variables are not crucial for business. Consumer Confidence Index is something that provides information about people's moods, but it is not something that a bank can influence during a marketing campaign. Products can and should be adapted to the situation on the markets, so the impact of these variables should be known, but from the point of view of the sales department they do not have the possibility to contact only those customers whose CCI is higher. As far as Outcome_cat is concerned, this is very important information, but again from the point of view of the organization - the new campaign cannot be limited to the previously contacted people only. That's why it was decided to build a model whose input consists only of personal features.

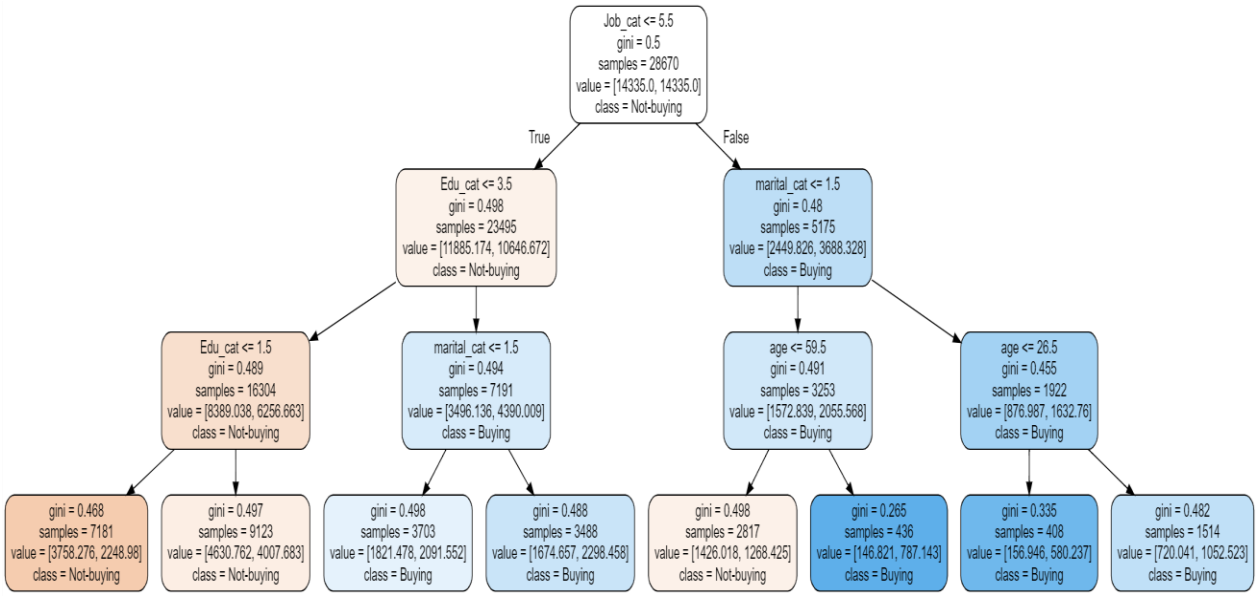

Figure 51 Decision Tree - Personal characteristics

Source: Author's own work

The decision tree built from personal attributes is shown in figure 51. The list of combinations of conditions for this tree can be found in table 28.

Leaf Node	**Conditions**	**Result**	**No. of samples**
1	Job_cat <= 5.5 Edu_cat <= 3.5 Edu_cat <= 1.5	Non-buying	7181
2	Job_cat <= 5.5 Edu_cat <= 3.5 Edu_cat > 1.5	Non-Buying	9123
3	Job_cat <= 5.5 Edu_cat > 3.5 Marital_cat <= 1.5	Buying	3703
4	Job_cat <= 5.5 Edu_cat > 3.5 Marital_cat > 1.5	Buying	3488

5	Job_cat > 5.5 Marital_cat <= 1.5 Age <= 59.5	Non-Buying	2817
6	Job_cat > 5.5 Marital_cat <= 1.5 Age > 59.5	Buying	436
7	Job_cat > 5.5 Marital_cat > 1.5 Age <= 26.5	Buying	408
8	Job_cat > 5.5 Marital_cat > 1.5 Age <= 26.5	Buying	1514

Table 28 Tree 3 splitting conditions

Figure 52 shows importance of personal features. In this case, we can see that apart from the marital_cat variable, all variables have a significant share in the classification of observations. Thus, using only personal attributes, we can see that bank deposits sold better among well-educated customers who were performing such professions as: administration, blue-collar, technician, services or management (left side of the tree) and among married people older than 59.5 years old or singles and divorced under 26.5 years old performing such professions as: retired, Entrepreneur, self-employed, housemaid, unemployed, student (right side of the tree). This kind of analysis of the interdependence of personal factors may prove to be helpful in the sales process.

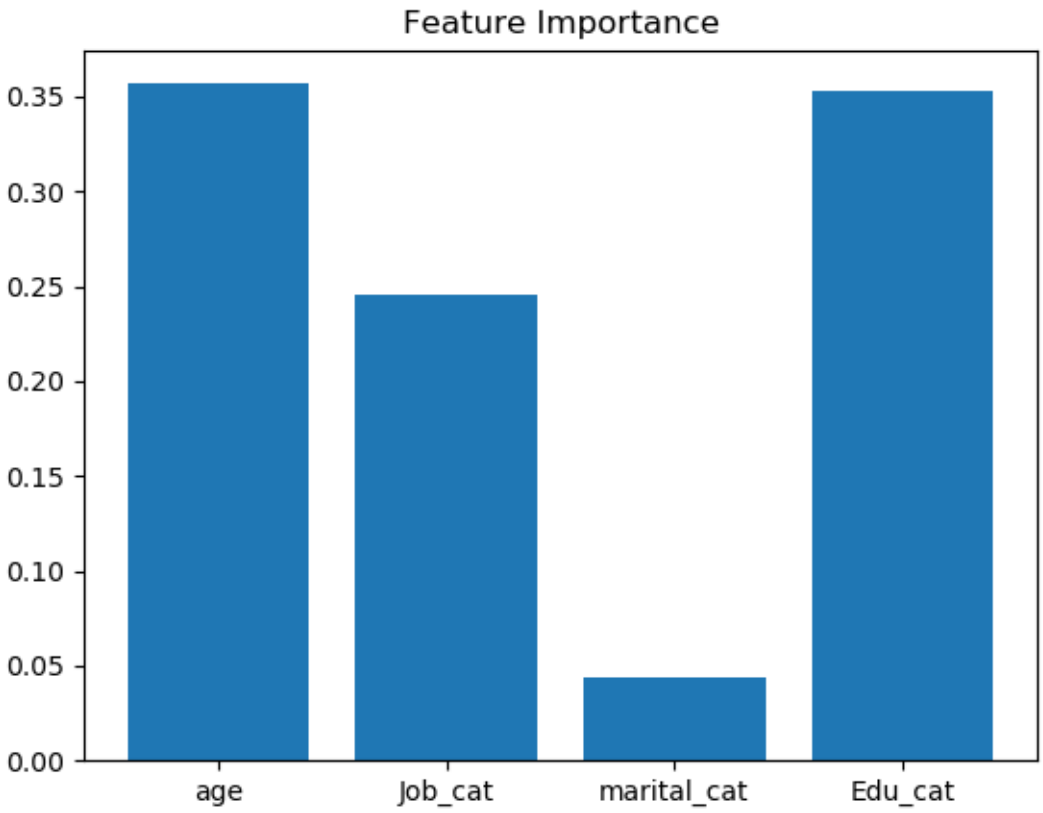

Figure 52 Tree 3 importance

Source: Author's own work

5. Results and Discussion

The next stage of the CRISP-DM methodology is to assess the predictive capabilities of previously developed models. For this purpose, some of the observations from the original data set were separated during the construction of the model and will now be used to measure its performance. This is the moment that determines whether the model behaves satisfactorily and can be approved for deployment or should be rebuilt.

5.1 Evaluation methods

The predictive abilities of the model should be verified at several levels in order to assess their behavior in different situations. The evaluation of logistic regression models and decision trees will be assessed using the following measures:

- Accuracy
- Confusion Matrix
- Area under the ROC curve (AUC)

5.2 Models evaluation

The model evaluation benchmark is derived from 25% of the data that were separated before the models were created, and then target variable predictions were created based on their input. The purpose of model evaluation is to determine how correct these predictions were. The first model to be evaluated will be logistic regression. The first criterion for evaluation of this model will be accuracy, which is an overall assessment of the accuracy of the model The logistic regression model achieved an overall accuracy score of 76.57%, which meant that more than 3/4 of the observations were classified successfully. The next evaluation criterion is the model's classification matrix. Here the precision and recall values for each class are evaluated. Table 29 presents confusion matrix for the logistic regression model. As we can see for class 1, precision, which is a ratio of true positive to total predicted positive is 27%. The recall value, i.e. the ratio of true positive to total actual positive, means how much true positive our model has captured, and takes the result 66%. This means that the model predicts more False Positive than False negative.

Class	**Precision**	**Recall**
0	95%	78%
1	27%	66%

Table 29 Logistic regression confusion matrix

The next criterion that we will use to evaluate the model is the Area under the ROC curve. For the logistic regression model it is shown in figure 53. It shows the true positive and false positive rates. The total AUC value was 0.72 and its shape increases quite sharply in the initial phase.

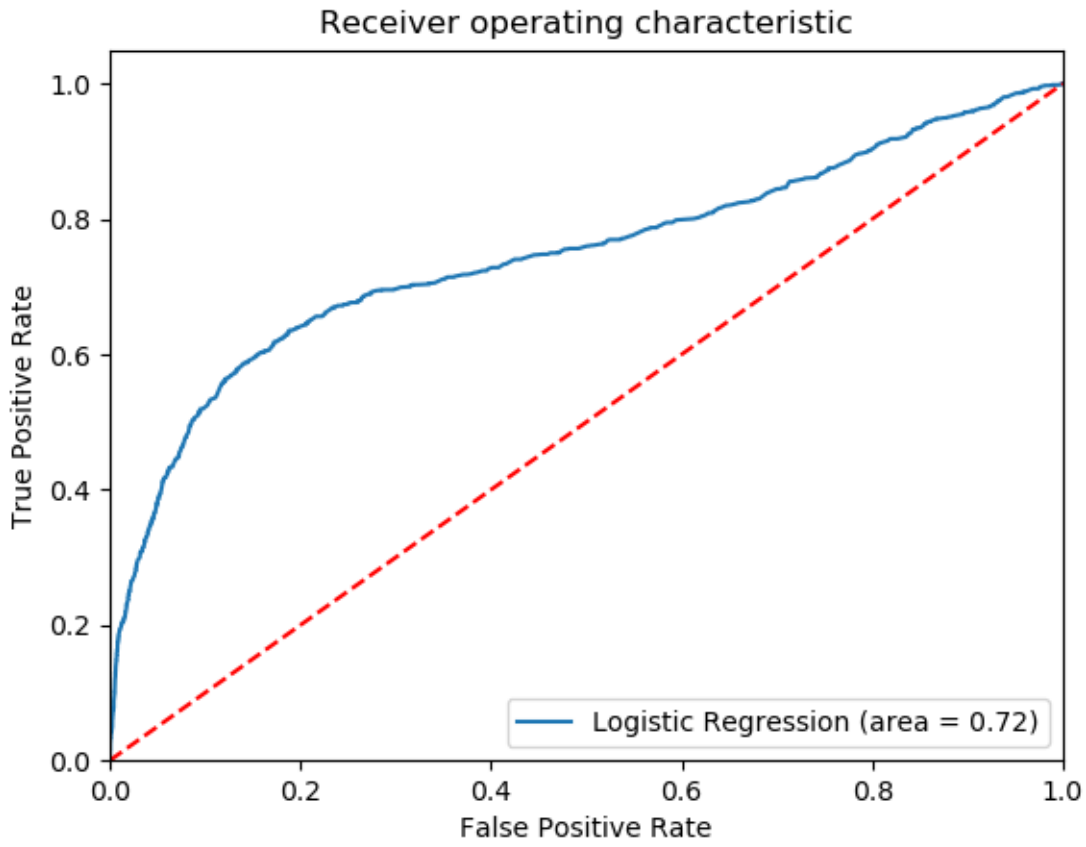

Figure 53 Logistic regression – AUC

Source: Author's own work

Subsequently, the decision trees have to be evaluated. As we can see in table 30, the trees of the full data set and the 3 best features achieved a very similar result, slightly different from each other. The overall accuracy score is lower for a tree using only personal features.

Tree	**Accuracy**
Full data set	88.53%
3 best features	88.46%
Personal features	65.87%

Table 30 Decision trees accuracy

Table 31 shows a comparison of precision and recall for each tree built during modelling. As we can see, the precision of the prediction is quite high, but for the target variable class = 1 their recall value is lower than for the logistic regression model. As we can see, full set and 3 best

features have almost identical balance, while a tree created from personal variables turned out to have better recall ratio and lower precision ratio for class = 1.

	Class	Precision	Recall
Full data set tree	0	92%	95%
	1	48%	37%
3 best features	0	92%	95%
	1	48%	37%
Personal features	0	91%	68%
	1	16%	49%

Table 31 Decision trees - confusion matrix

Figures 54 and 55 show the AUC for decision trees. As we can see, their result is clearly lower than the logistical regression.

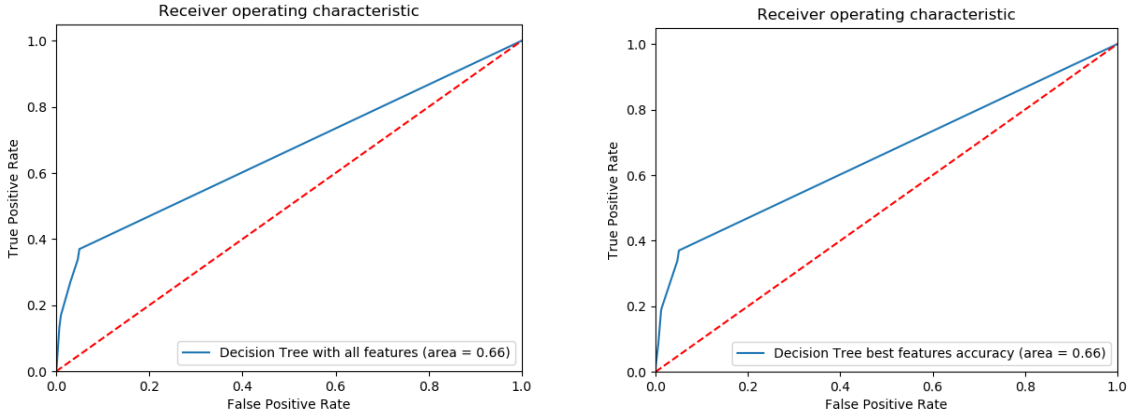

Figure 54 AUC for tree 1 & 2

Source: Author's own work

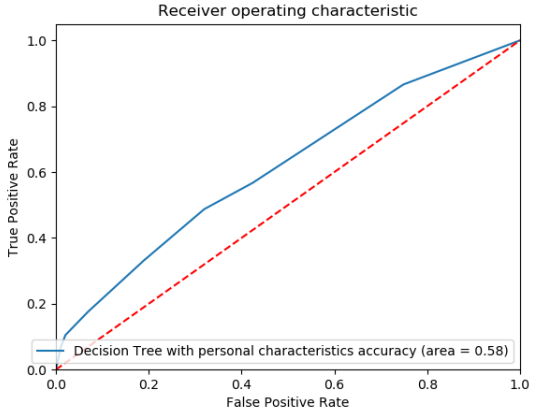

Figure 55 AUC for tree 3

Source: Author's own work

Given the overall specification and the aim of the study, in its current form, despite the lower overall accuracy value, the results of logistical regression are more relevant to the campaign than the decision trees. The tree models are characterized by much more precise predictions and a better prediction of class = 0, while the logistic regression, despite the poorer behavior for class = 0, found the observations of class = 1 better. Given that the sales department during the marketing campaign has the appropriate means to make contacts, as well as the specification of the business, seems to be more interested in better capturing customers who will make purchases than in general precision, the logistic regression model seems to be more appropriate here.

The ability to view logistic regression coefficients allows us to interpret the influence of individual factors on the final result. Therefore, we are able to compare the findings that our model presents to the characteristics of individual factors mentioned in the theoretical part. Table 32 compares the influence of the final model parameters with the expected result.

Model parameter	**Model impact**	**Expected impact**
Customer experience	Positive	Either positive or negative
Interest rate (eurlibor3m)	Negative	Positive
Contact	Negative	Cannot be determined
Inflation (Consumer Price Index)	Positive	Positive
Consumer Confidence Index	Positive	Positive
Marital_single	Positive	Cannot be determined
Education level	Positive	Positive
Personal loan	Negative	Negative

Table 32 Results impact comparison

The first parameter, customer experience, represents what kind of experience the customer has with the product or more generally the company through the learning process. The experience can be both positive and negative, however, in our model, the positive impact of this variable means that the experience gained by customers was more likely to be positive.

The second parameter, interest rate from a theoretical point of view is a significant factor that has a positive impact on the willingness to invest. When the interest rate is higher, people more often decide to invest, including the purchase of bank deposits because it is more beneficial to them. Nevertheless, our model unexpectedly presents a negative relationship.

Inflation, which is very strongly linked to the interest rate, should also have a positive relationship with deposit sales. As inflation rises, those funds that have been accumulated so far lose their purchasing power. Therefore, in situations of rising inflation, the demand for investments should also grow. In this case, the impact of the variable in the model is compatible with expectations.

The Consumer Confidence Index is a variable that describes consumer moods about the economy and financial situation. Better consumer mood implies less concern about saving and increasing consumption and investment. The result achieved in the model is in line with expectations.

Customers interested in purchasing bank deposits as an investment good are those who have sufficient financial resources. Research on the correlation between the level of education and the earnings of individuals clearly shows that a higher level of education usually translates into higher earnings. Therefore, we can expect that a higher level of education will have a positive impact on deposit sales. In this case, the model has behaved as expected.

The direct impact of some of the variables in this specific situation and referring to this specific product was not unambiguously anticipated and is marked in the table as 'cannot be determined'.

6. Conclusion

The presented thesis has a theoretical understanding of the terminology of data analysis and big data. Moreover, it lists types, methods and applications of various forms of data analysis. The third chapter, consisting of a review of the literature, brings closer the theoretical foundations of the predictive analytics environment. It also presents data analysis methodologies and techniques that can be used for these needs. The third chapter also discusses in detail the set of data that will be analyzed using other analysts' opinions on the influence of given factors.

Chapter four is a practical application of the methodology and techniques described in the previous chapter. The aim of the ongoing research was to find factors that will help optimize the sale of bank deposits during a marketing campaign and/or reduce its cost, i.e. make more efficient use of the resources available. For the purpose of building the model, also due to the interpretative possibilities, was chosen: Logistic regression and decision trees. Chapter 4 also assumes optimization of the model after its creation. The 5th chapter consists of evaluation of the achieved results in terms of overall effectiveness and, above all, meeting the assumed objectives of the analysis. At this point it has also been confirmed that all, apart from one parameter of the final model, are behaving in a manner consistent with the expectations of the previously selected theoretical analysis.

The current research has shown that following the methodology of the CRISP-DM process we are able to perform a full-scale data analysis. The process, which was initially presented in the theoretical part, was then used in the hands-on part. Its phases included, in successive stages, understanding why business needs analysis, understanding the data to be analyzed, then adjusting and preparation, and finally modelling and evaluation.

Each of the created models was characterized by different characteristics. Although the best overall accuracy result was achieved with the use of the decision tree model, its strength was based on effective prediction of the negative class. The logistic regression model, despite its lower overall predictive effectiveness, was characterized by a higher ability to identify the positive class, expressed in the highest value of the recall value, thus allowing for the discovery of those people the business would like to reach. Ultimately, carrying out this type of predictive analysis will allow to identify the people who are most interested in the contact, allowing to increase the effectiveness of the operation. Therefore, conducting future campaigns will be optimized to achieve a result similar to the current one, with less costs resulting from the reduction in the number of contacts or using the same amount of available materials to achieve higher sales. It also represents a kind of profit for customers, assuming that they are aware customers. This will allow those customers who are potentially interested in purchasing this product to be contacted, thus avoiding making unwanted contacts.

Big data analytics, which is currently on the boom stage, provides a great opportunity to gain knowledge from the inconspicuous and underestimated bits of information. There are many reasons for this development, however, it is essentially the evolution and the current technology that has pushed the opportunities of data analysis to unprecedented boundaries. Data flows nowadays come from many different sources, sometimes requiring specific infrastructure and considerable effort to bring them to a structured form. There are many applications of current data analysis or so called data science, incredibly overwhelming what statistics were used for half a century ago. Thanks to these capabilities we are able to monitor many aspects of everyday life, increasing the effectiveness of our actions as well as helping to understand the world around us. It is worth noting the ethical aspect of this evolution. Personally, I believe that in a world of general digitization, it is extremely important that the algorithms surrounding us ubiquitously focus on seeking relationships that lead to an overall improvement in the lives of societies, and not simply shaping perfect customers among the people. Optimizing the business or even the economy is a great success, however, data analysis may also help in areas such as poverty reduction, natural disasters or global pandemics. I think that the last and most important link that contributes to the development of data analysis is the analytic community. They have been and are the driving force behind the implementation of various models and methods to the problems of the world around us and ultimately help to solve them. An example of such activity could be the activity of so-called hackathons, i.e. events aimed at solving a predetermined problem, usually in a relatively short period of time. One of these hackathons is Covid-19 challenge organized by MIT[5]. It aims to find solutions which most affect health care and citizens during a global pandemic. With such initiatives, the analytic community can serve laudable purposes and create a truly valuable analysis.

Works Cited

Abbott, D. (2014). *Applied Predictive Analytics.* Indianapolis: John Wiley & Sons, Inc.

Bahga, A., & Madisetti, V. (2019). *Big Data Analytics: A Hands-On Approach.* Arshdeep Bahga & Vijay Madisetti.

[5] https://covid19challenge.mit.edu/

Bellman, R. (1957). *Dynamic Programming.* Princeton University Press.

Cristianini, N., & Taylor, J. S. (2013). *An introduction to Support Vector Machines and other kernel based learning methods.* Cambridge: Cambridge University Press.

Dornbusch, R. (2011). *Macroeconomics.* University of Washington.

EMMI. (2019, 11 24). Retrieved from https://www.emmi-benchmarks.eu/

Falcidieno, B., Pienovi, C., & Spagnuolo, M. (2005). Descriptive modeling and prescriptive modeling in spatial data handling. *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, 122-135.

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data .* Hampshire: Palgrave MACMILLAN.

Glauber, D. E. (1964). Multicollinearity in Regression Analysis. *Sloan School of Management Massachusetts Institute of Technology*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to Statistical Learning.* London: Springler.

Khandelwal, R. (2020, 01 19). *Medium*. Retrieved from Toward data science: https://towardsdatascience.com/feature-selection-identifying-the-best-input-features-2ba9c95b5cab

Kotler, P., & Armstrong, G. (2016). *Principles of Marketing.* Harlow: Pearson Education Limited.

Kotu, V., & Deshpande, B. (2015). *Predictive Analytics and Data Mining.* Elsevier Science & Technology.

Kumar, B. (2011). Filter versus Wrapper Feature Subset Selection in. *International Journal of Computer Science and Information Technologies*, 1048-1053.

Larose, D. T., & Larose, D. C. (2015). *Data Mining and Predictive Analytics.* New Jersey: John Wiley & Sons, Inc.

Linoff, G. S., & Berry, M. J. (2011). *Data Mining Techniques.* New Jersey: John Wiley & Sons, Incorporated.

Morina, F., & Osmani, R. (2019). The impact of macroeconomic factors on the level of deposits in the banking sector,. *Journal of Accounting, Finance and Auditing Studies*, 16-29.

Müller, A., & Guido, S. (2017). *Introduction to Machine learning with Python.* Sebastopol: O'Reilly Media.

OECD. (2019, 11 24). Retrieved from https://data.oecd.org/emp/employment-rate.htm

Ohlhorst, F. J. (2013). *Turning Big Data into Big Money.* New Jersey: John Wiley & Sons, Inc.

Online Etymology Dictionary. (2019, 11 16). Retrieved from https://www.etymonline.com/word/analysis

Oracle. (2019, 11 16). Retrieved from Big Data: https://www.oracle.com/big-data/guide/what-is-big-data.html

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 559-572.

Pedregosa, F. a. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 2825-2830.

Phillips, A. W. (1958). The Relation between Unemployment and the Rate of Change of Money Wage. *Economica*, 283-299.

Pose, A. R., & Vassilis, T. (2006). Education and income inequality in the regions of European Union.

Sarkar, D. (2020, 01 18). *Medium*. Retrieved from Towards data science: https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63

Sauter, V. L. (2010). *Decision Support Systems for Business Intelligence.* New Jersey: John Wiley & Sons Inc.

Strawson, P. F. (1992). Analysis and Metaphysics. *Oxford: Oxford University Press*, 2.

Turhani, A., & Hoda, H. (2016). The Determinative Factors of Deposits Behavior in Banking System in Albania. *Academic Journal of Interdisciplinary Studies*, 246-256.

Zhang, A. (2017). *Data Analytics.*

List of figures

List of tables

Appendix

Variable	**Levels**	**Meaning**
Eurlibor3m_above_3	1;0	i > 3%; i <3%
Contact_cat	1;2;3	‘campaign’ = 1;=2;>2

CPI_cat	1;2;3;4;5	‘cons.price.idx’ < 93;<=93.5;<=94;<=94.5;>94.5
CCI_cat	1;2;3;4	‘cons.conf.idx’ <-45;<=-40;<=-35;>-35
Personal_loan	1;0	True;False
Mortgage	1;0	True;False
Edu_cat	1;2;3;4	'education' = basic.4y | basic.6y | basic.9y; high.school; professional.course; University degree

Appendix 1 Logistic regression categorical variables coded

Variable	**Levels**	**Meaning**
Job_cat	1;2;3;4;5;6;7;8;9;10;11	1 = ‘admin.’; 2 = ’blue-collar’; 3 = ’technician; 4 = ‘services’; 5 = ‘management’; 6 = ‘retired’; 7 = ‘entrepreneur’; 8 = ‘self-employed’; 9 = ‘housemaid; 10 = ‘unemployed’ 11 = ‘student’
Contact_cat	1;2;3	‘campaign’ = 1;=2;>2
CPI_cat	1;2;3;4;5	‘cons.price.idx’ < 93;<=93.5;<=94;<=94.5;>94.5
CCI_cat	1;2;3;4	‘cons.conf.idx’ <-45;<=-40;<=-35;>-35
Personal_loan	1;0	True;False
Mortgage	1;0	True;False
Edu_cat	1;2;3;4	'education' = basic.4y | basic.6y | basic.9y; high.school; professional.course; University degree
Outcome_cat	1;2;3	1 = ‘nonexistent’; 2 = ‘failure’; 3 = ‘success’
Rate_cat	1;2;3;4;5;6	'euribor3m' < 1; <= 2; <=3; <=4; <=5;>5

Appendix 2 Decision tree categorical variables coded

Logistic Regression Model

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import statsmodels.api as sm
from scipy.stats import norm
import matplotlib.pyplot as plt
from sklearn.metrics import classification_report
from sklearn.metrics import roc_auc_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_curve
import seaborn as sns

#Display settings for PyCharm
desired_width=420
pd.set_option('display.width', desired_width)
np.set_printoptions(linewidth=desired_width)
pd.set_option('display.max_columns',50)
pd.set_option('display.float_format', lambda x: '%.4f' % x)

#import data
df_bank = pd.read_csv("bank-additional-full.csv")
df_bank =
df_bank.drop(columns=['default','contact','day_of_week','duration','pdays','previous','nr.
employed','month'])

#delete 'unkown' observations
df_bank = df_bank[df_bank.job != 'unknown']
df_bank = df_bank[df_bank.marital != 'unknown']
df_bank = df_bank[df_bank.education != 'unknown']
df_bank = df_bank[df_bank.housing != 'unknown']
df_bank = df_bank[df_bank.loan != 'unknown']
df_bank = df_bank[df_bank.education !='illiterate']

#delete emp.var.rate column
df_bank = df_bank.drop(columns=['emp.var.rate'])


#correct skewness - log+1 function
age_log = np.log(df_bank.age + 1)
print(age_log.skew())
print(df_bank.age.skew())
df_bank = df_bank.assign(age_log=pd.Series(np.log(df_bank.age + 1)))

#create categorical features
df_bank = pd.get_dummies(df_bank, columns=['poutcome'], drop_first=True)
df_bank = df_bank.assign(eurlibor3m_above_3 = np.where(df_bank['euribor3m'] > 3, 1,0))
df_bank = df_bank.assign(Contact_cat= np.where(df_bank['campaign'] == 1, 1,
```

```
                                    np.where(df_bank['campaign'] == 2, 2, 3)) )
df_bank = df_bank.assign(CPI_cat =np.where(df_bank['cons.price.idx'] < 93, 1,
                                    np.where(df_bank['cons.price.idx'] <= 93.5, 2,
                                        np.where(df_bank['cons.price.idx'] <= 94, 3,
                                            np.where(df_bank['cons.price.idx'] <= 94.5, 4, 5)))))
df_bank = df_bank.assign(CCI_cat = np.where(df_bank['cons.conf.idx'] <-45, 1,
                                    np.where(df_bank['cons.conf.idx']<= -40, 2,
                                        np.where(df_bank['cons.conf.idx'] <= -35, 3, 4 ))))


#encode to bool
df_bank = df_bank.assign(subscribed = np.where(df_bank['y'] == 'yes', 1, 0) )
df_bank = df_bank.assign(personal_loan = np.where(df_bank['loan'] =='yes', 1, 0))
df_bank = df_bank.assign(mortgage = np.where(df_bank['housing'] == 'yes', 1, 0))

#Change categorical to bool / ordinal
df_bank= pd.get_dummies(df_bank, columns=['marital','job'], drop_first=True)
df_bank = df_bank.assign(Edu_cat =np.where(df_bank['education'] == 'basic.4y', 1,
                                    np.where(df_bank['education'] == 'basic.6y', 1,
                                        np.where(df_bank['education'] == 'basic.9y', 1,
                                            np.where(df_bank['education'] == 'high.school', 2,
                                                np.where(df_bank['education'] ==
'professional.course', 3,4))))) )

df_bank = df_bank.drop(columns=['age','cons.price.idx','cons.conf.idx',
'y','loan','housing','campaign','euribor3m','education'])
df_bank2= df_bank.loc[:, df_bank.columns != 'subscribed']
#assign input variables
y = df_bank['subscribed']
X = df_bank.loc[:, df_bank.columns != 'subscribed']
print(X.head(n=10))
#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#Logistic regression
LR_Model = LogisticRegression(random_state=0 , class_weight='balanced')
LR_Model.fit(X_train,y_train)
LR_Pred = LR_Model.predict(X_test)
print("Logistic regression accuracy =",accuracy_score(y_test, LR_Pred)*100,"%")
print(X.columns)
print(LR_Model.coef_)
print(LR_Model.intercept_)

# --------------------
```

```
def logit_pvalue(model, x):
    """ Calculate z-scores for scikit-learn LogisticRegression.
    parameters:
        model: fitted sklearn.linear_model.LogisticRegression with intercept and large C
        x:     matrix on which the model was fit
    This function uses asymtptics for maximum likelihood estimates.
    """
    p = model.predict_proba(x)
    n = len(p)
    m = len(model.coef_[0]) + 1
    coefs = np.concatenate([model.intercept_, model.coef_[0]])
    x_full = np.matrix(np.insert(np.array(x), 0, 1, axis = 1))
    ans = np.zeros((m, m))
    for i in range(n):
        ans = ans + np.dot(np.transpose(x_full[i, :]), x_full[i, :]) * p[i,1] * p[i, 0]
    vcov = np.linalg.inv(np.matrix(ans))
    se = np.sqrt(np.diag(vcov))
    t =  coefs/se
    p = (1 - norm.cdf(abs(t))) * 2
    return p

# test p-values
model = LR_Model
np.set_printoptions(suppress=True)
df1 = pd.DataFrame(logit_pvalue(model, X_train),columns=['P_value']).astype(float)
df2 = pd.DataFrame(X.columns)
df3 = pd.DataFrame(np.array(["Const"]))
df_valid = df3.append(df2 ,ignore_index=True, sort=False)
df_valid = pd.concat([df_valid, df1],axis=1)
print(df_valid)
print(df_valid[df_valid.P_value < 0.05])


#Re-build model

df_bank = df_bank.drop(columns=['age_log',
'mortgage','marital_married','job_entrepreneur','job_management','job_self-
employed','job_technician','job_unemployed'])

#verify model after rebuilding
y = df_bank['subscribed']
X = df_bank.loc[:, df_bank.columns != 'subscribed']
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)
LR_Model = LogisticRegression(random_state=0 , class_weight='balanced')
LR_Model.fit(X_train,y_train)
LR_Pred = LR_Model.predict(X_test)
```

```
print("Logistic regression accuracy =",accuracy_score(y_test, LR_Pred)*100,"%")
print(X.columns)
print(LR_Model.coef_)
print(LR_Model.intercept_)
model = LR_Model
np.set_printoptions(suppress=True)
df1 = pd.DataFrame(logit_pvalue(model, X_train),columns=['P_value']).astype(float)
df2 = pd.DataFrame(X.columns)
df3 = pd.DataFrame(np.array(["Const"]))
df_valid = df3.append(df2 ,ignore_index=True, sort=False)
df_valid = pd.concat([df_valid, df1],axis=1)
print(df_valid[df_valid.P_value < 0.05])

cnf_matrix = confusion_matrix(y_test, LR_Pred)
print(cnf_matrix)

print(classification_report(y_test, LR_Pred))

logit_roc_auc = roc_auc_score(y_test, LR_Model.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, LR_Model.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

Appendix 3 Logistic regression code

Decision Tree Model

```
import pandas as pd
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV
import matplotlib.pyplot as plt
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_auc_score
```

```
from sklearn.metrics import roc_curve
from sklearn.tree import export_graphviz


#Display settings for PyCharm
desired_width=420
pd.set_option('display.width', desired_width)
np.set_printoptions(linewidth=desired_width)
pd.set_option('display.max_columns',50)

#import data
df_bank = pd.read_csv("bank-additional-full.csv")
df_bank =
df_bank.drop(columns=['default','contact','day_of_week','duration','pdays','previous','nr.
employed','month'])

#delete 'unkown' observations
df_bank = df_bank[df_bank.job != 'unknown']
df_bank = df_bank[df_bank.marital != 'unknown']
df_bank = df_bank[df_bank.education != 'unknown']
df_bank = df_bank[df_bank.housing != 'unknown']
df_bank = df_bank[df_bank.loan != 'unknown']
df_bank = df_bank[df_bank.education !='illiterate']

#delete emp.var.rate column
df_bank = df_bank.drop(columns=['emp.var.rate'])

#correct skewness - log+1 function
'''
age_log = np.log(df_bank.age + 1)
print(age_log.skew())
print(df_bank.age.skew())
df_bank = df_bank.assign(age_log=pd.Series(np.log(df_bank.age + 1)))
'''
#Create numerical categories
df_bank = df_bank.assign(Job_cat= np.where(df_bank['job'] == 'admin.',1,
                                 np.where(df_bank['job']=='blue-collar',2,
                                     np.where(df_bank['job']=='technician', 3,
                                         np.where(df_bank['job']=='services', 4,
                                             np.where(df_bank['job']=='management', 5,
                                                 np.where(df_bank['job']=='retired', 6,

np.where(df_bank['job']=='entrepreneur', 7,
                                                                     np.where(df_bank['job']=='self-
employed', 8,
```

```
np.where(df_bank['job']=='housemaid', 9,

np.where(df_bank['job']=='unemployed', 10,

np.where(df_bank['job']=='student', 11, 0))) ) ) ) ) ) ) ) ) )
df_bank = df_bank.assign(Contact_cat= np.where(df_bank['campaign'] == 1, 1,
                                    np.where(df_bank['campaign'] == 2, 2, 3)) )
df_bank = df_bank.assign(CPI_cat =np.where(df_bank['cons.price.idx'] < 93, 1,
                                  np.where(df_bank['cons.price.idx'] <= 93.5, 2,
                                      np.where(df_bank['cons.price.idx'] <= 94, 3,
                                          np.where(df_bank['cons.price.idx'] <= 94.5, 4, 5)))))
df_bank = df_bank.assign(CCI_cat = np.where(df_bank['cons.conf.idx'] <-45, 1,
                                   np.where(df_bank['cons.conf.idx']<= -40, 2,
                                       np.where(df_bank['cons.conf.idx'] <= -35, 3, 4 ))) )
df_bank = df_bank.assign(marital_cat = np.where(df_bank['marital'] == 'married',1,
                                    np.where(df_bank['marital']=='single',2,
                                        np.where(df_bank['marital'] == 'divorced', 3,0))))
df_bank = df_bank.assign(Edu_cat =np.where(df_bank['education'] == 'basic.4y', 1,
                                  np.where(df_bank['education'] == 'basic.6y', 1,
                                      np.where(df_bank['education'] == 'basic.9y', 1,
                                          np.where(df_bank['education'] == 'high.school', 2,
                                              np.where(df_bank['education'] ==
'professional.course', 3,4))))) )
df_bank = df_bank.assign(Outcome_cat = np.where(df_bank['poutcome'] =='nonexistent',1,
                                    np.where(df_bank['poutcome'] == 'failure',2,
                                        np.where(df_bank['poutcome'] == 'success', 3,0))) )
df_bank = df_bank.assign(rate_cat = np.where(df_bank['euribor3m'] <= 1, 1,
                                   np.where(df_bank['euribor3m'] <= 2,2,
                                       np.where(df_bank['euribor3m'] <=3,3,
                                           np.where(df_bank['euribor3m'] <=4,4,
                                               np.where(df_bank['euribor3m'] <=5,5,6)))
)))
#encode to bool
df_bank = df_bank.assign(subscribed = np.where(df_bank['y'] == 'yes', 1, 0) )
df_bank = df_bank.assign(personal = np.where(df_bank['loan'] =='yes', 1, 0))
df_bank = df_bank.assign(mortgage = np.where(df_bank['housing'] == 'yes', 1, 0))

df_bank = df_bank.drop(columns=['job','marital',
'education','loan','housing','campaign','euribor3m','poutcome','cons.price.idx','cons.conf
.idx','y'])

print(df_bank.sample(n=10))

#assign input variables
y = df_bank['subscribed']
X = df_bank.loc[:, df_bank.columns != 'subscribed']
```

```
#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#build model
clf = DecisionTreeClassifier(random_state=0, class_weight='balanced',
max_depth=3,max_features='sqrt',min_samples_leaf=10,min_samples_split=2)
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Decision Tree with all features accuracy =",accuracy_score(y_test,
y_pred)*100,"%")

#Find value of splitt importance
importances = clf.feature_importances_
plt.figure()
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances)
plt.xticks(range(X.shape[1]), X.columns)
plt.show()

#Print confusion matrix
cnf_matrix = confusion_matrix(y_test, y_pred)
print(cnf_matrix)
print(classification_report(y_test, y_pred))

#Print AUC value
DT_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Decision Tree with all features (area = %0.2f)' % DT_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()

#Export graphiz
dot_data = export_graphviz(clf,out_file='Tree1.dot',feature_names=X.columns
                    ,filled=True,rounded=True, class_names=['Not-buying','Buying'])

#Decision Tree with personal characteristics
X = df_bank[['age','Job_cat','marital_cat','Edu_cat']]
```

```
#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#build model
clf = DecisionTreeClassifier(random_state=0, class_weight='balanced',
max_depth=3,max_features='sqrt',min_samples_leaf=10,min_samples_split=2)
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Decision Tree with personal characteristics accuracy =",accuracy_score(y_test,
y_pred)*100,"%")

#Find value of splitt importance
importances = clf.feature_importances_
plt.figure()
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances)
plt.xticks(range(X.shape[1]), X.columns)
plt.show()

#Print confusion matrix
cnf_matrix = confusion_matrix(y_test, y_pred)
print(cnf_matrix)
print(classification_report(y_test, y_pred))

#Print AUC value
DT_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Decision Tree with personal characteristics accuracy (area =
%0.2f)' % DT_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()

#Export graphiz
dot_data = export_graphviz(clf,out_file='Tree2.dot',feature_names=X.columns
                           ,filled=True,rounded=True, class_names=['Not-buying','Buying'])

#Best-important features
X = df_bank[['age','CCI_cat','Outcome_cat']]
print(X.head)
```

```
#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#build model
clf = DecisionTreeClassifier(random_state=0, class_weight='balanced',
max_depth=3,max_features='sqrt',min_samples_leaf=10,min_samples_split=2)
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Decision Tree best features accuracy =",accuracy_score(y_test, y_pred)*100,"%")
importances = clf.feature_importances_

#Show importance
plt.figure()
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances)
plt.xticks(range(X.shape[1]), X.columns)
plt.show()

#Print confusion matrix
cnf_matrix = confusion_matrix(y_test, y_pred)
print(cnf_matrix)
print(classification_report(y_test, y_pred))

#Print AUC
DT_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Decision Tree best features accuracy (area = %0.2f)' %
DT_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()

#Export graphiz
dot_data = export_graphviz(clf,out_file='Tree3.dot',feature_names=X.columns
                    ,filled=True,rounded=True, class_names=['Not-buying','Buying'])

#Make DT with features selected in Logistic Regression
X =
df_bank[['CCI_cat','Outcome_cat','rate_cat','CPI_cat','personal','marital_cat','Job_cat',
```

```
'Edu_cat']]

#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#build model
clf = DecisionTreeClassifier(random_state=0, class_weight='balanced',
max_depth=3,max_features='sqrt',min_samples_leaf=10,min_samples_split=2)
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Decision Tree Log reg features accuracy =",accuracy_score(y_test,
y_pred)*100,"%")
importances = clf.feature_importances_

#Show which features are the most important
plt.figure()
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances)
plt.xticks(range(X.shape[1]), X.columns)
plt.show()

#Print confusion matrix
cnf_matrix = confusion_matrix(y_test, y_pred)
print(cnf_matrix)
print(classification_report(y_test, y_pred))

#Print AUC
DT_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Decision Tree Log reg features (area = %0.2f)' % DT_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()

#Export graphiz
dot_data = export_graphviz(clf,out_file='Tree4.dot',feature_names=X.columns
                    ,filled=True,rounded=True, class_names=['Not-buying','Buying'])


#Make DT with client record data
```

```
X = df_bank[['Outcome_cat','personal','mortgage']]

#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#build model
clf = DecisionTreeClassifier(random_state=0, class_weight='balanced',
max_depth=3,max_features='sqrt',min_samples_leaf=10,min_samples_split=2)
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Decision Tree client features accuracy =",accuracy_score(y_test, y_pred)*100,"%")
importances = clf.feature_importances_

#Show which features are the most important
plt.figure()
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances)
plt.xticks(range(X.shape[1]), X.columns)
plt.show()

#Print confusion matrix
cnf_matrix = confusion_matrix(y_test, y_pred)
print(cnf_matrix)
print(classification_report(y_test, y_pred))

#Print AUC
DT_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Decision Tree client features (area = %0.2f)' % DT_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()

#Export graphiz
dot_data = export_graphviz(clf,out_file='Tree5.dot',feature_names=X.columns
                           ,filled=True,rounded=True, class_names=['Not-buying','Buying'])

#Make DT with client record data
X = df_bank[['CPI_cat','CCI_cat','rate_cat']]
```

```
#Split values to train and test
X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.25, random_state=1)

#build model
clf = DecisionTreeClassifier(random_state=0, class_weight='balanced',
max_depth=3,max_features='sqrt',min_samples_leaf=10,min_samples_split=2)
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Decision Tree macroeconomical features accuracy =",accuracy_score(y_test,
y_pred)*100,"%")
importances = clf.feature_importances_

#Show which features are the most important
plt.figure()
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances)
plt.xticks(range(X.shape[1]), X.columns)
plt.show()

#Print confusion matrix
cnf_matrix = confusion_matrix(y_test, y_pred)
print(cnf_matrix)
print(classification_report(y_test, y_pred))

#Print AUC
DT_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Decision Tree macroeconomical features (area = %0.2f)' %
DT_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()

#Export graphiz
dot_data = export_graphviz(clf,out_file='Tree6.dot',feature_names=X.columns
                           ,filled=True,rounded=True, class_names=['Not-buying','Buying'])
```

Appendix 4 Decision Tree code