Open University of Cyprus

School of Economics and Management

Master's Degree

Enterprise Risk Management

Master Thesis

Big Data Analytics in Risk Management: Credit Risk Assessment and Evaluation using Machine Learning Algorithms

Platias Christos

Supervisor
Dr Pantelis Ipsilantis

May 2019

Open University of Cyprus

School of Economics and Management

Master's Degree

Enterprise Risk Management

Master Thesis

Big Data Analytics in Risk Management: Credit Risk Assessment and Evaluation using Machine Learning Algorithms

Platias Christos

Supervisor
Dr Pantelis Ipsilantis

This postgraduate dissertation was submitted for partial fulfillment of the

requirements for obtaining a postgraduate degree on Enterprise Risk Management from the School of Economics Science and Management of the Open University of Cyprus

May 2019

To my Son Peter…

Table of Contents

Table of Figures

Abstract

The scope of this Master Thesis is to introduce and analyze the concept of credit scoring and credit risk evaluation through an extensive description and implementation of an end to end classification modelling process in a both analytical and business manner. An open source large dataset which includes almost 1 million loan applications is used to construct and implement 3 different classification models in order to give answers to the following questions; What is the process of constructing a classification model for credit risk scoring? What incremental value does a classification model for predicting the probability of default adds to a financial institution? And finally, which classification algorithm suits better for the purpose of credit scoring calculation and risk assessment? Extensive literature is introduced in order to construct a proper theoretical and modelling framework for the final implementation of the 3 different models and their results as well as a detailed comparison in terms of accuracy and sensitivity are presented.

Chapter 1
Introduction

The scope of this Master Thesis was to introduce some of the most common concepts of credit risk management through an end to end classification modelling process for the correct estimation of the probability an individual may default on their obligations towards a financial institution. In the first part of this Thesis, extensive literature regarding big data evolution and implementation of machine learning and statistical techniques for the calculation of credit scoring were introduced as a forerunner, to develop a specific framework derived from the literature for the implementation of classification methods in credit scoring calculation. Initially, several definitions of the concept of Big Data are introduced in order to properly define the core aspects of the field. Madden, defines Big Data as: *"data that's too big, too fast, or too hard for existing tools to process.* (Madden, 2012). Provost and Fawcett, (2013), define Big Data as *"datasets that are too large for traditional data-processing systems and that therefore require new technologies.* One of the most significant definition is the one that Gartner proposed. He defines Big Data with reference to what he calls "the 3 V's": Volume, Velocity, Variety. Big data practice in Enterprises and the concept of credit risk and credit scoring are extensively introduced in order to define the framework of the modelling part. Kenton defines credit risk is as "*the probable risk of loss resulting from a borrower's failure to repay a loan or meet contractual obligations*." (Kenton, 2018). Abdou and Pointon, argue that credit evaluation is one of the most crucial processes in banks' credit management decisions. This process includes collecting, analyzing and classifying different credit elements and variables to assess the credit decisions. Credit scoring calculation methods are thoroughly presented through detailed review of the existing literature. Calculation methods are distinct into statistical and Machine Learning techniques and the most significant attributes of the most used classification algorithms are explained.

The second part of this Master Thesis contains the methodological framework of the analysis, research problem and research questions, modelling process and the results of the different classification algorithms implemented in order to define the most proper classification algorithm for the calculation of credit scoring. A large, open-source dataset with almost 1 million observations and 74 different characteristics of applicants who have asked for funding from Lending Club between 2007 and 2015 was used for the analysis. A very extensive presentation of the quality of all 74 different variables is performed through descriptive statistics. This is a very interesting but also important part of an analytical procedure in order to better understand the business problem and define the most appropriate strategy. Three well known and common in practice classification algorithms were used in order to predict the credit score of an applicant. Logistic Regression, Decision Trees and K Nearest Neighbor classification algorithms have been tested in terms of accuracy, sensitivity and specificity. The results and comparisons of those algorithms are introduced in detail through 3 of the most common evaluation metrics such as Area Under Curve, Confusion Matrixes and ROC. SMOTE resampling technique is implemented in the initial dataset in order to correct the specificity and sensitivity of the model in predicting the true defaulters. Lending Club dataset is highly imbalanced towards minority class observations and the initial samples failed to properly train the algorithms towards the maximization of sensitivity. In the last part of this Master Thesis, the research questions are answered according to the results of the modelling process and further investigation and research is proposed in terms of whether other ML algorithms than those 3 implemented may result in greater predictive power.

Chapter 2

Big Data: The Management revolution

Data; a very fascinating and mainstream term nowadays which dynamically influences our lives and has deeply changed the way we operate, think and live. Data have their roots in the ancient years when the first scientists collected their knowledge in ancient papyrus in order to communicate their exceptional findings and expertise to the next generations. During centuries, huge amounts of information were collected and stored in libraries, notebooks, papers, in any conservative or not conservative way. Data is everywhere; from a notebook where we used to write our notes in the classroom our first years in school, to huge data warehouses where major companies and global organizations, countries and governments collect information and intelligence in order to better understand their needs and facilitate their functionality. In the information era, enormous amounts of data have become available on hand to decision makers. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data.

In the first chapters of this Master Thesis, the Author will go through a detailed presentation of Big Data, their definition, concepts and implementation, in order to better understand their power in interpreting, communicating and deriving knowledge but also, in generalized decision making.

2.1 Big Data: Definitions and Concepts

Although Big Data is a trending buzzword in both academia and the industry, its meaning is still shrouded by much conceptual vagueness. The term is used to describe a wide range of concepts: from the technological ability to store, aggregate, and process data, to the cultural shift that is pervasively invading business and society, both drowning in information overload. The lack of a formal definition has led research to evolve into multiple and inconsistent paths. Furthermore, the existing ambiguity among researchers and practitioners undermines an efficient development of the subject. (De Mauro, Greco, & Grimaldi, 2015)

Big Data possesses multiple and diverse nuances of meaning, all of which have the right to exist. By analyzing the most significant occurrences of this term in both academic and business literature we have identified four key themes to which Big Data refers: Information, Technologies, Methods and Impact. We can reasonably assert that the vast majority of references to Big Data encompass one of the four themes listed above. (De Mauro, Greco, & Grimaldi, 2015)

Hadi and Shnain (2015), mention that the term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005. Magoulas used the term Big Data, in order to define a great amount of data that traditional data management techniques cannot manage and process due to their complexity and size (Hadi & Shnain, 2015). In the same paper, we can find another definition of Big Data, provided by Madden (2012). Madden, define the Big Data as: *"data that's too big, too fast, or too hard for existing tools to process."* Hadi explains that, "Too big" means that organizations must increasingly deal with enormous collections of data that come from click streams, transaction histories and sensors, "Too fast" means that not only the data is extremely large, but additionally, should be processed in a very quick manner and finally, "Too hard", means that such data may not be easily processed by existing tools, with a need of even some more analysis, not suited to existing tools (Hadi & Shnain, 2015).

Provost and Fawcett, (2013), define Big Data as *"datasets that are too large for traditional data-processing systems and that therefore require new technologies"* with names like Hadoop and MongoDB. Ehrenberg (2012) notes that when he first used the term "big data" in lower case in 2009 to label a new ventures fund, the term "implied tools for managing large amounts of data and applications for extracting value from that data". Cloudera CEO Mike Olson describes Big Data as complex data at volume, but he admits to not really liking the term Big (Power, 2014). Gandomi and Haider (2015), with a sense of humor, argue that *“big data … probably originated in lunch-table conversations at Silicon Graphics Inc. in the mid-1990s”*. Google, describes Big Data through their implementation on data analytics and their ability to discover meaning from conservative data (Ward & Barker, 2013).

According to Ward, (2013), big data is predominantly associated with two ideas: data storage and data analysis. Despite the fact that interest in big data is quiet new and the same characterization may be applied on their implementation in enterprises’ decision making, these concepts are far from new and have long lineages (Ward & Barker, 2013). This, therefore, raises the question as to how big data is typically different from current conventional data processing techniques. In order to get an answer on this question, one should need to look no further than the term big data. “Big” implies significance, complexity and challenge. Unfortunately, the term “big” also invites quantification and therein lies the difficulty in furnishing a definition (Gandomi & Haider, 2015).

Hadi argues that in order to understand Big Data, we should consider the ways data can actually supports real-life profitable or beneficial outcomes (Hadi & Shnain, 2015). Enterprises all over the world have recently begun to exploit the power of Big Data. Many companies have been experimenting with techniques that allow them to collect massive amounts of data in order to determine whether hidden patterns exist within that data that might be an early indication of an important change. Data might show, for example, that customer buying patterns are changing or that new factors affecting the business must be considered (Hadi & Shnain, 2015).

Big Data is important because they provide organizations with the power to gather, store, manage, and manipulate vast amounts of data, at the right time, with the right speed, in order to gain the right insights (Gandomi & Haider, 2015).

Amongst the most known and cited definitions of Big Data is that, introduced by Gartner in 2001 Gartner proposed a threefold definition encompassing the “three Vs” of Big Data: Volume, Velocity, Variety. This definition has later been reinforced by the definition of 5 V’s of Big Data, which is commonly accepted nowadays. In the 3V’s of Gartner, Veracity and Value have been added to the existing Volume, Velocity and Variety model **(Figure 1):**

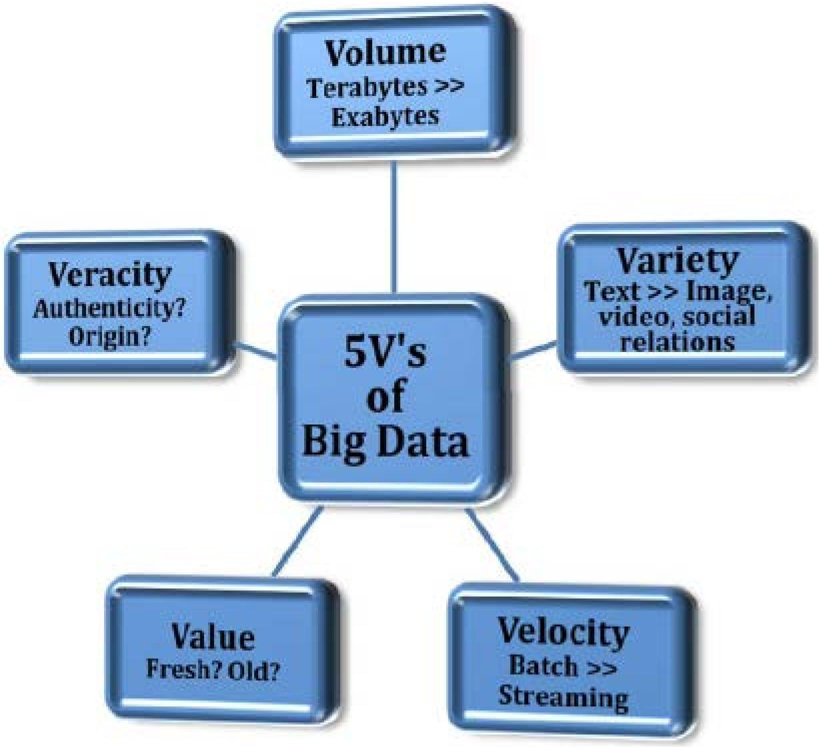


Figure 1: The 5 V's of Big Data

- Volume refers to the quantity of data gathered by a company. This data must be used further to gain valuable and important knowledge. The volume of data defines their characterization. If the volume o is very large then it is actually considered as a ‘Big Data’. This means whether a particular data can actually be considered as a Big Data or not, is dependent only upon their volume.
- Velocity: refers to the time in which Big Data can be processed. Some activities are very important and need immediate responses, which is why fast processing maximizes efficiency. For time-sensitive processes such fraud detection, Big Data flows must be analyzed and used as they stream into the organizations in order to maximize the value of the information (Hadi & Shnain, 2015).

- Variety: According to Gandomi (2015), variety, refers to the type of data that Big Data can comprise. This data may be structured or unstructured. There is a great variety in the types of data. Big data consists in any type, including structured and unstructured data such as text, sensor data, audio, video, click streams, log files and so on (Hadi & Shnain, 2015).
- Value: refers to the important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis/hypothesis. Data value will depend on the events or processes they represent such as stochastic, probabilistic, regular or random. Depending on this the requirements may be imposed to collect all data, store for longer period (for some possible event of interest), etc. In this respect data value is closely related to the data volume and variety.
- Veracity: refers to the degree in which a leader trusts information in order to make a decision. Therefore, finding the right correlations in Big Data is very important for the business future. However, as one in three business leaders do not trust the information used to reach decisions, generating trust in Big Data presents a huge challenge as the number and type of sources grows.

2.2 Big data evolution

Everyone knows that the Internet has changed how businesses operate, governments function, and people live. But a new, less visible technological trend is just as transformative: “big data.” Big data starts with the fact that there is a lot more information floating around these days than ever before, and it is being put to extraordinary new uses. Big data is distinct from the Internet, although the Web makes it much easier to collect and share data. Big data is about more than just communication: the idea is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts.

Bernard Marr (2015), on his article published in World Economic Forum Website in 2015, demonstrates a brief review of the historical evolution of Big Data, starting from Ancient times and evolving, through nowadays, with the great impact of Internet and big data warehouses in the ways we run companies and society.

Marr, mentions that the earliest examples we have of humans storing and analyzing data are the tally sticks almost 20.000 years ago. The Ishango Bone which was discovered in 1960 in what is where now Uganda located, is thought to be one of the earliest pieces of evidence of prehistoric data storage (Marr, 2015). In 2400 BC, the abacus – the first dedicated device constructed specifically for performing calculations – comes into use in Babylon. Marr argues that the same time the first libraries made their appearance, indicating our first attempts at mass data storage (Marr, 2015).

In the third century BC, the Library of Alexandria was believed to house the sum of human knowledge. Today, there is enough information in the world to give every person alive 320 times as much of it as historians think was stored in Alexandria's entire collection -- an estimated 1,200 Exabyte' worth. If all this information were placed on CDs and they were stacked up, the CDs would form five separate piles that would all reach to the moon.

On 100 BC, the first known computer machine has been produced, presumably by Greek scientists. The Antikythera Mechanism **(Figure 2),** the earliest discovered mechanical computer, was supplied with a "CPU" consisted of 30 interlocking bronze gears and it is thought to have been designed for astrological purposes and tracking the cycle of Olympic Games (Marr, 2015).

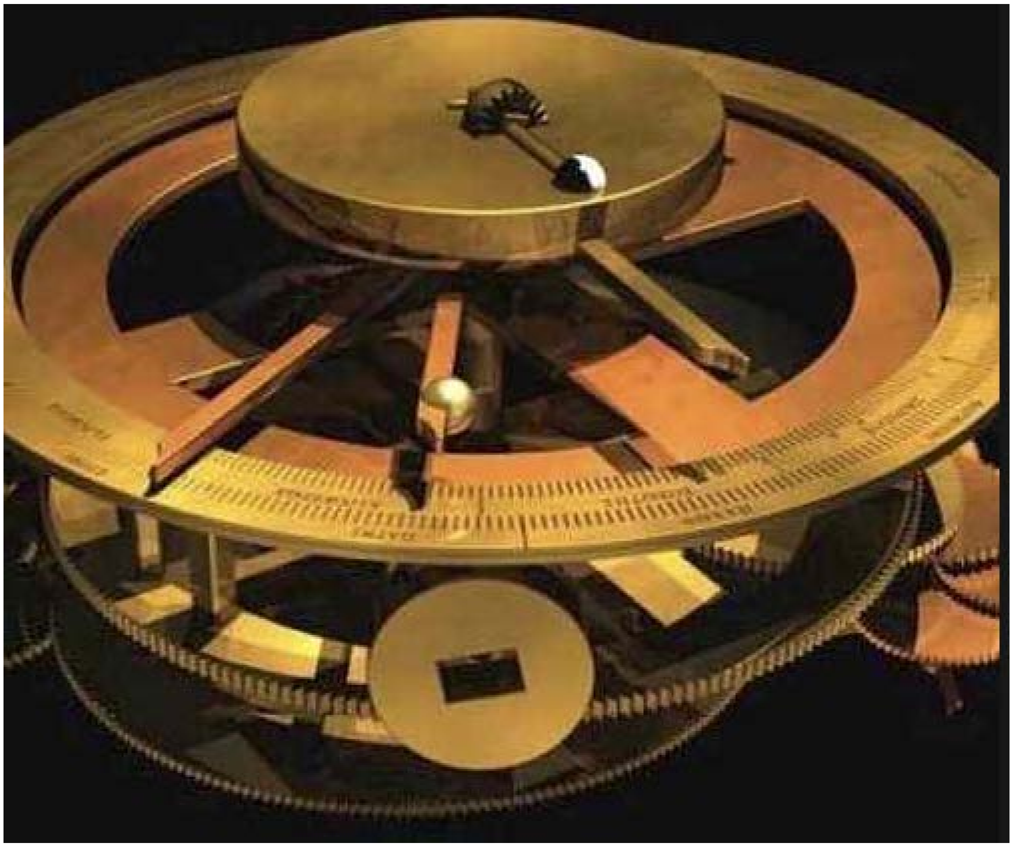


Figure 2: A demonstration of the Antikythera Mechanism

Marr continuous his historical review with the emerge of statistics, the *"calculation expert infant"* of Mathematics, which influenced the most the advances in data collection and analysis. It was the year 1661 when John Graunt carried out the first recorded experiment in statistical data analysis. By recording information about mortality, he theorized that he could design an early warning system for the bubonic plague ravaging Europe (Marr, 2015).

Almost 200 years after Graunt's work, Herman Hollerith, a young engineer employed by the US Census Bureau, produces what will become known as the Hollerith Tabulating Machine **(Figure 3).** Marr mentions that in 1880, the Bureau was unable to process all data collected in the 1880 census. Bureau's estimations were that it would take almost 8 years to crunch all the data collected and it was predicted that the data generated by the 1890 census would take another 10 years, meaning that it would be impossible to assess those data before they would be outdated by the 1900 census. Hollerith, used punch cards, in order to reduce 10 years' work to three months and achieves his place in history as the father of modern automated computation. The company he founds will go on to become known as IBM (Marr, 2015).

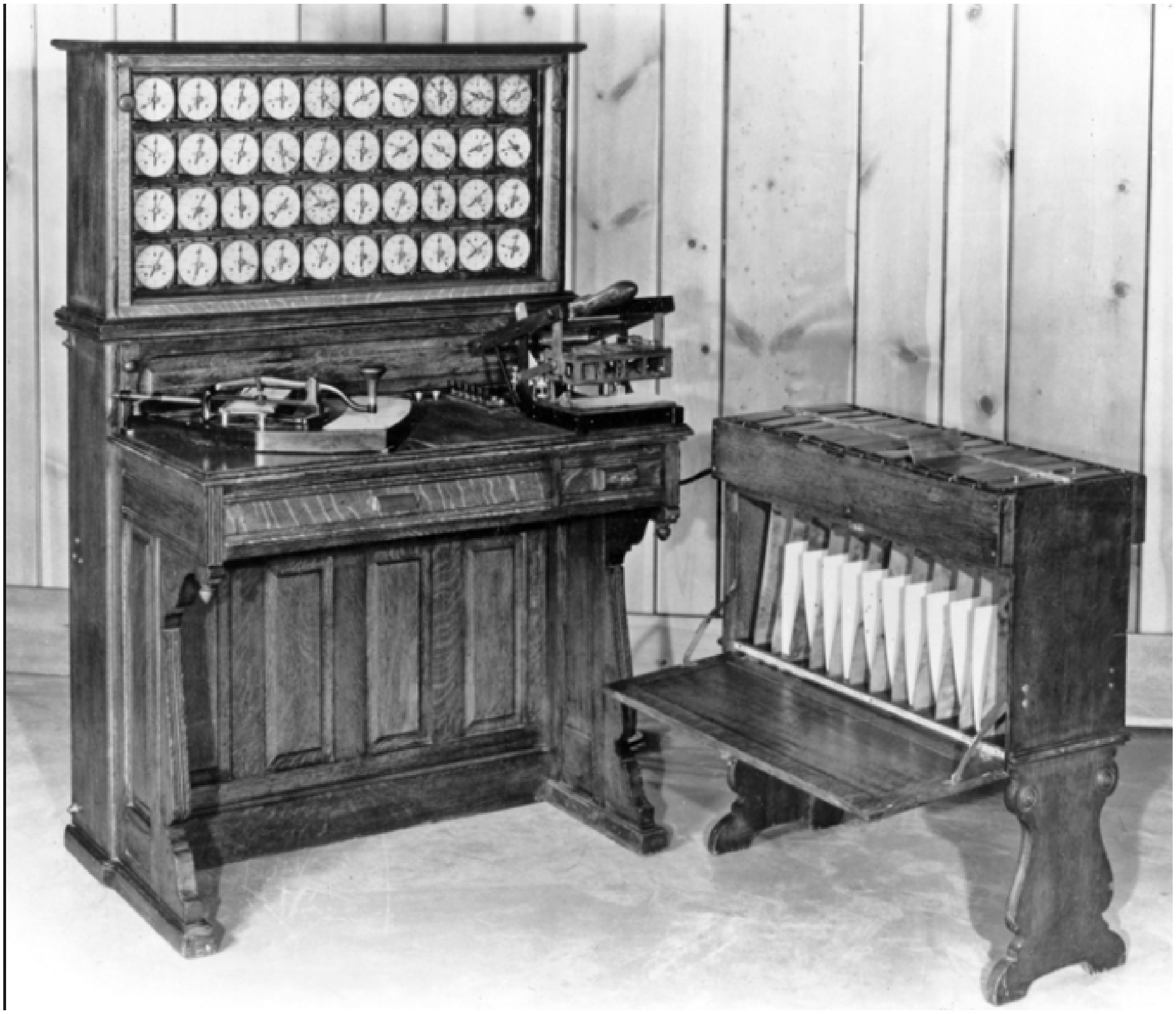


Figure 3: Hollerith's Tabulating Machine

Nicola Tesla, in 1926, in an interview at Colliers Magazine, stated that "when wireless technology is "perfectly applied the whole Earth will be converted into a huge brain, which in fact it is, all things being particles of a real and rhythmic whole". Tesla did now that he was describing one of the greatest inventions of human kind, the World Wide Web. Two years later, on 1928, Fritz Pfleumer, a German-Austrian engineer, invents a method of storing information magnetically on tape. Marr says that the principles Pfleumer used in his invention are still in use today, with the vast majority of digital data being stored magnetically on computer hard disks (Marr, 2015).

IBM's research, during 60's and 70's should be regarded as one of the biggest contributions in the beginning of Business Intelligence. IBM engineer William C Dersch presented the Shoebox Machine at the 1962 World Fair. It could interpret numbers and sixteen words spoken in the English language into digital information. This was one of the first steps towards speech recognition algorithms (Marr, 2015). IBM mathematician Edgar F Codd presented his framework for a "relational database". This framework, is used by many modern data services nowadays, in order to store information in a hierarchical format, which could be accessed by anyone who knows what they are looking for (Marr, 2015).

Marr mentions the birth of World Wide Web in 1991 as a benchmark in the evolution of Big Data. Eight years later, and the term Big Data appears in Visually Exploring Gigabyte Datasets in Real Time, published by the Association for Computing Machinery (Marr, 2015). The paper went on to quote computing and automation pioneer Richard W Hamming as saying: *"The purpose of computing is insight, not numbers."*

In 2008, world's servers were capable of storing 9.6 zettabytes (9.6 trillion of gigabytes) of information – almost 12 gigabyte per person living in Earth. On 2009, the average US Company stored over 200 terabytes of data according to a report of McKinsey Digital (Manyika & Chui, 2011). In 2010, Google chairman Eric Schmidt, on a conference in Digital Innovations, supported that *"as much data is now being created every two days, as was created from the beginning of human civilization to the year 2003"*, (Marr, 2015).

From the review represented, one may think that explosion of data is relatively new. The amount of digital data expands so quickly -- doubling around every three years -- Today, less than two percent of all stored information is non digital (Marr, 2015).

Given this massive scale, it is tempting to understand big data solely in terms of size. But that would be misleading. We can learn from a large body of information things that we could not comprehend when we used only smaller amounts. This kind of data is being put to incredible new uses with the assistance of inexpensive computer memory, powerful processors, smart algorithms, clever software, and math that borrows from basic statistics. Instead of trying to “teach” a computer how to do things, such as drive a car or translate between languages, which artificial-intelligence experts have tried unsuccessfully to do for decades, the new approach is to feed enough data into a computer so that it can infer the probability that, say, a traffic light is green and not red or that, in a certain context, *lumière* is a more appropriate substitute for “light” than *léger*. (Cukier & Schoenberger, 2013)

2.3 Big data in enterprises

The constantly changing environment in the digital economy has challenged traditional economic and business concepts. Huge volumes of user-generated data are transferred and analyzed within and across different sectors, gradually increasing the markets’ dependency on precise and timely information services. George et al., arguing towards the importance of very small “unimportant in a naive eye thing” in the stability of an enterprise (George, Haas, & Pentland, 2014). For example, they refer to the impact of a mere tweet from a trusted source that can cause losses or profits of billions of dollars and a chain reaction in the press, social networks and blogs. This situation makes information goods even more difficult to value as they have a catalytic impact on real-time decision-making. In contrast, entrepreneurs and innovators have taken aggregate open and public data as well as self-quantification and exhaust data to create new products and services that have the power to transform industries. In private and public spheres, Big Data sourced from mobile technologies and banking services such as digital/mobile money when combined with existing ‘low tech’ services such as water or electricity can transform societies and communities. There is little doubt that over the next decade it will change the landscape of social and economic policy and research. (George, Haas, & Pentland, 2014)

Chapter 3

Credit Risk: Definitions and concepts

3.1 The concept of credit risk

In order to describe the concept of credit risk in an easier, more friendly to a naïve reader's way, let's introduce a simple, everyday example:

Let's take the case that someone you may have known at school or a friend from the past, turns to you and asks you to lend them some money with a predefined certain interest rate, let's say 20%. The amount you are asked to borrow is not a trivial one to buy for example a ticket to the theater but an amount that if your "friend" will not pay you back as promised, you would be left significantly out of money.

So, what would you do? Would you lend the person the money? They may not repay you. Therefore, maybe it would be better to refuse. On the other hand, an interest rate of 20% is not considered as an indifferent proposal and if refuse, you may lose out on a possible profitable opportunity. The crux of the decision is whether the individual will keep the promise to repay or defaults. The best-case scenario is that the loan is repaid (with the aforementioned interest). The undesirable outcome that you wish to avoid is that the individual fails to repay the loan or, in the parlance of credit, defaults.

Note how the example raises all sorts of issues. If you knew the individual better, you might be more inclined to go with the lending decision (that is, if you knew the person's circumstances and their ability to repay). The past experience of others who have lent money to the individual might be useful to know. You may also wish to compare the individual to others who have borrowed money in a similar situation. As a result, you may be able to obtain a statistical estimate of the likelihood that the individual will repay you (or, equivalently, will default on the loan). Your views as to whether you would be wise to lend the money to this acquaintance might change if the individual produced a guarantee to support the loan, or some collateral (that is, something you could call upon if the individual were unable or unwilling to meet the obligation). Whatever your thoughts, the decision requires you to make a judgement on the uncertain future outcome. This might take the form of a gut feeling (or what professionals would term expert judgement), or you might be able to rely on a formal assessment model.

The previous example aimed to introduce the concept of credit risk and the parameters the lender should take into consideration towards the decision to proceed to the proposed deal or not.

There are many formal definitions of credit risk proposed in the literature. According to Kenton for example, credit risk is "*the probable risk of loss resulting from a borrower's failure to repay a loan or meet contractual obligations*." (Kenton, 2018). Traditionally, it refers to the risk that a lender may not receive the owed principal and interest, which results in an interruption of cash flows and increased costs for collection. A more detailed definition of the concept of credit risk is given by Brown & Moles in their book "Credit Risk Management" (Brown & Moles, 2016).

"Credit risk can be defined as 'the potential that a contractual party will fail to meet its obligations in accordance with the agreed terms'."

Credit risk is also variously referred to as **default risk, performance risk** or **counterparty risk.** These all fundamentally refer to the same thing: the impact of credit effects on a firm's transactions (Brown & Moles, 2016).

There are three characteristics that define credit risk:

1. Exposure (to a party that may possibly default or suffer an adverse change in its ability to perform).
2. The likelihood that this party will default on its obligations (the default probability).
3. The recovery rate (that is, how much can be retrieved if a default takes place

Although it is impossible to know exactly who will default on obligations, properly assessing and managing credit risk can lessen the severity of loss.

But how risky are such situations for enterprises and organizations and how severe may be the consequences of a possible default or failure to meet contractual obligations?

“The company is fundamentally sound. The balance sheet is strong. Our financial liquidity has never been stronger… My personal belief is that Enron stock is an incredible bargain at current prices and we will look back a couple of years from now and see the great opportunity that we currently have. “

Enron Chairman Kenneth Lay on 26 September 2001 in an online chat with employees, as reported by Reuters. Less than three months later the company filed for bankruptcy.

As the quotation above indicates, a transaction may expose the buyer to unforeseen outcomes. While buying stock can be considered inherently risky, in that you are taking a chance on the firm’s performance, most firms have similar problems: to gain orders, they may need to lend money to customers through granting credit terms on sales or they may ask other firms to undertake work on their behalf; or they may place surplus cash on deposit with a financial institution. Hence, as a result of transactions of various kinds, credit risk and credit risk management are key issues for most firms.

3.2 Credit Scoring: How to evaluate credit risk?

Credit evaluation is one of the most crucial processes in banks’ credit management decisions (Abdou & Pointon, 2011). This process includes collecting, analyzing and classifying different credit elements and variables to assess the credit decisions. The quality of bank loans is the key determinant of competition, survival and profitability. So, the main target of banks’ decision making relies on the early identification of the quality of loans in terms of the probability to default. One of the most important kits, to classify a bank’s customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring.

According to Hand & Jacka (Hand & Jacka, 1998) credit scoring is *“the process (by financial institutions) of modelling creditworthiness”*. In other words, credit scoring refers to the process that a lender applies in order to evaluate the “credit health” of the borrower in terms of repaying their obligations. Thus, credit scoring refers to the set of decision models and techniques that aid lenders in granting consumer credit by assessing the risk of lending to different consumers (Bellotti & Crook, 2009). Mester (Mester, 1997) describes credit scoring as *“a method of evaluating the credit risk of loan applications”*.

A set of decision models and their underlying techniques that aid lenders in the granting of consumer credit (Gup & Kolari, 2005). Thomas (Thomas L. C., 2000), indicates that those techniques tent to “decide” who will get credit, how much credit and most important, what operational strategies will enhance the probability and the amount of profit of the borrowers to the lenders.

Anderson (Anderson, 2007), introduces the concept of credit scoring as the use of statistical models to “translate” all available data into numerical indicators that are going to be used to grade credit decisions. He pinpoints the importance for organizations and enterprises to forecasting financial risk and introduces the concepts of credit and behavioral scoring as the applications of financial risk forecasting towards consumer lending (Thomas L. C., 2000). According to the author, credit and behavioral scoring are the techniques that aid financial institutions to decide whether or not to grand loans and credits to consumers.

A most recent definition of credit scoring is given by Louzada et al. (Louzada, Anderson, & Guilherme, 2016) in their research on classification techniques applied in credit scoring. The authors define credit scoring as *“a numerical expression based on a level analysis of customer credit worthiness”*, a helpful tool for assessment and prevention of default risk, an important method in credit risk evaluation, and an active research area in financial risk management.

3.2.1 Credit Scoring: Historical review

Thomas (Thomas L. C., 2000) indicates that credit scoring is a way to recognize and discriminate between the different groups of a population when the different characteristics of the groups are not clear enough. The concept of discrimination between different groups of a population was initially introduced in statistics by Fisher back in 1936. He sought to differentiate between two varieties of iris (a small species of flowering plants with showy flowers) by measurements of the physical size of the plants and to differentiate the origins of the skulls using their physical measurements (Fisher, 1986). 5 years later, David Durant introduces the idea of using Fisher's techniques and conceptual framework to discriminate between good and bad loans. (Durant, 1941). Durant's work was a research project and was never implemented for any predictive purposes. In the same period some of the main financial and credit institutes were facing difficulties in managing their credit obligations. Decision on whether to give a loan or not had been made judgmentally by credit analysts for many years. However, those analysts had been draft into military service and thus, there was a severe shortage of people with that kind of expertise for many years. This shortage in expertise led the firms to get their analysts right down their subjective rules that used to decide to whom to give loans; that was one of the first examples of expert systems (Johnson, 1992).

So, how exactly did credit analysts finally decided on giving or rejecting a loan? As Thomas describes (Thomas L. C., 2000), their decision was influenced from the so called "5C's":

1. The **C**haracter of the person (do I know the person or their family?)
2. The **C**apital (how much is being asked for?)
3. The **C**ollateral (what is the applicant willing to put up from their own resources)
4. The **C**apacity (what is their repaying ability?)
5. The **C**ondition (what are the conditions in the market?)

The 5 C's, clearly reveal in a very comprehensive way that the final decision on whether a loan should be granted or not, was a subject of personal judgment; In fact, credit analyst was the one and only responsible to decide, with all the severe consequences of any mistaken judgment.

It did not take long after the war for some innovators to start thinking of the use of statistical methods and models in lending decisions (Thomas L. C., 2000). The first consultancy firm in the US was formed in the early 1950's. Their clients where mainly finance houses retailers and mail order firms. The most significant innovations of those days that influenced the new wave towards statistical modeling for credit decision making was the arrival of credit cards in 1960. The number of people applying for a credit card every day made it impossible both in economic and manpower terms to do anything but automate the lending decision (Thomas L. C., 2000). When credit organizations started using credit scoring in order to evaluate the applications, they found out that it was also a much better predictor that any judgmental scheme to evaluate the probability to default; in fact, default rate dropped almost 50% during the next decade (Myers & Forgy, 1963).

The success of credit scoring in credit cards led banks to start using credit scoring also for their other products like personal loans. In 1980'S financial institutions started using scoring for home loans and small business loans. In the 1990's the significant growth in direct marketing led to the use of scorecards to evaluate the efficiency of firms marketing campaigns (Abdou & Pointon, 2011). Also, in 1980's logistic regression and linear programming, two of the most useful techniques in credit risk evaluation where introduced. Advances in computing power allowed other techniques to be tried to build credit scoring models.

Nowadays, the emphasis is on trying to change the objectives from minimizing the probability that a customer will default on a certain product (a loan for example), to looking at how the firm can maximize the profit from that particular customer (Thomas L. C., 2000). The original idea of estimating the probability to default has been replaced by estimations of response rate (how probable is a consumer to respond to a direct mailing of a new product), usage rate (how likely is a consumer to use a product), retention rate (how likely is the consumer to keep using the product after the initial offering period) and of course, debt management; (if the consumer starts to become delinquent on the loan how successful are various approaches to prevent default).

3.2.2 Credit scoring today

As mentioned above, credit scoring methods are widely used to estimate and to minimize credit risk. Mail order companies, advertising companies, banks and other financial institutions use these methods to score their clients, applicants and potential customers. There is effort to precise all procedures used to estimate and decrease credit risk.

Nevertheless, applications of credit scoring have been widely used in different fields, including a comparison between different statistical techniques used in prediction purposes and classification problems (Abdou & Pointon, 2011).

Both the U.S. Federal Home Loan Mortgage Corporation and the U.S. Federal National Mortgage Corporation have encouraged mortgage lenders to use credit scoring which should provide consistency across underwriters. Also, the international banks supervision appeals to precise banks internal assessments: The Basel Committee on Banking Supervision is an international organization which formulates broad supervisory standards and guidelines for banks. It encourages convergence toward common approaches and common standards. The Committee’s members come from Belgium, Canada, France, Germany, Italy, Japan, Luxembourg, 10 the Netherlands, Spain, Sweden, Switzerland, United Kingdom and United States. In 1988, the Committee decided to introduce a capital measurement system (the Basel Capital Accord). This framework has been progressively introduced not only in member countries but also in other countries with active international banks. In June 1999, the Committee issued a proposal for a New Capital Adequacy Framework to replace the 1988 Accord (http://www.bis.org). The proposed capital framework consists of three pillars: 1. minimum capital requirements, 2. supervisory review of internal assessment process and capital adequacy, 3. effective use of disclosure to strengthen market discipline.

3.3 How is credit scoring calculated?

3.3.1 General framework

As we described before, credit scoring is a method of evaluating the credit risk of loan applications. Lim and Sohn (Lim & Sohn, 2007) in their article on deploying a cluster based scoring model, propose that the categorization of good and bad credit is of fundamental importance, and is the core objective of a credit scoring model. The need of an appropriate classification technique is thus evident. But what determines the categorization of a new applicant? How exactly do the corresponding metrics calculate?

Analysts use historical data and statistical techniques in order to produce a "score" that a bank can use to rank its loan applicants or borrowers in terms of risk (Mester, 1997). In other words, credit scoring *"tries to isolate the effects of various applicant characteristics on delinquencies and defaults"* (Mester, 1997). To build a scoring model, or "scorecard," developers analyze historical data on the performance of previously made loans to determine which borrower characteristics are useful in predicting whether the loan performed well. A well-designed model should give a higher percentage of high scores to borrowers whose loans will perform well and a higher percentage of low scores to borrowers whose loans won't perform well. Information on borrowers is obtained from their loan applications and from credit bureaus (Thomas L. C., 2000).

From the review of literature, characteristics such as gender, age, marital status, having a telephone, educational level, occupation, time at present address and having a credit card are widely used in building scoring models (Sustersic, Mramor, & Zupan, 2009) (Hand, Sohn, & Kim, Optimal bipartite scorecards, 2005) (Lee & Chen, 2005). Time at present job, loan amount, loan duration, house owner, monthly income, bank accounts, having a car, mortgage, purpose of loan, guarantees and others have been also used in building the scoring models (Sarlija, Bensic, & Bohacek, 2004), (Lee & Chen, 2005). In some cases, the list of variables has been extended to include spouse personal information, such as age, salary, bank account and others. Of course, more variables are less frequently used in building scoring models, such as television area code, weeks since the last county court judgement, worst account status, time in employments, time with bank and others (Banasik & Crook, 2007) (Bellotti & Crook, 2009).

3.3.2 Credit Scoring Calculation

In this section, we will describe in a more detailed, statistical way, how credit scoring works in order to understand the concept before diving deeper into classification algorithms and machine learning processes.

As we have already discussed, one of the main aims of the credit risk manager is to analyze different dimensions and aspects of an applicant’s profile in order to assess whether or not this individual would be on time at their obligations towards the financial institute or not. In other words, discriminate applicants in two big, mutual exclusive categories; those lenders that will pay their loans in time, and those that default on their loan within given time. However, due to the fact that no manager is able to “predict” a future outcome (unless he is a wizard), he does not know the type of a client beforehand and needs to decide whether to give a loan based on a set of variables provided by the client themselves (application data), third party data providers (credit agencies’ data) or historical behavior of the customer (data on previously taken loans) (Herasymovych, 2018).

Usually, the lender has a sample of loans that were given to clients and matured, thus letting the manager observe characteristics of borrowers and corresponding outcomes of the credit-granting decision. Thus, the problem can be described as a simple classification task; Will the borrower pay their obligations or not. More details about classification problems will be discussed in the following chapters of this Master Thesis.

Let’s denote the vector of characteristics of a loan application (the variables used for the classification) as X and the outcome as a binary variable Y, which is 1 if the loan is bad (the borrower defaults on his obligations) and 0 if the loan is good (the payment is made in time). Then a variety of classification algorithms (e.g. logistic regression, decision trees, etc.) can be applied to predict the outcome variable or estimate the probability of the loan being bad. In other words, the algorithm will apply a value between 0 and 1 on the probability that a borrower would fail to pay their loan:

$$P(Bad\ Loan\,|\,Vector\ of\ Variables) = P(Y=1\,|\,X) = \hat{y}$$

The estimated probability is transformed afterward, recalibrating to a more comprehensible range (Thomas, Crook, & Edelman, 2017) and possibly adjusting for company's policy objectives and rules resulting in a credit score (s^{CS}) for companies' applications. Then the decision maker applies an acceptance threshold (t^{AT}), a cutoff edge towards the decision to accept or reject the loan (if $s^{CS} \geq t^{AT}$ then the loan is accepted, otherwise it is rejected) (Thomas, Crook, & Edelman, 2017).

The predictive performance of the model is then assessed comparing predicted outcomes to actual ones for a test dataset independent from the one the model was trained on. For example, let's say that the variables that credit analyst use in order to predict the probability of default are annual income, marital status, other current loans that the applicant may have, occupational category, annual loan rates from the National Bank, country's economy status (from A: Excellent economy to F: Economy under surveillance), amount of the loan applied and a dummy variable indicating whether the applicant lives in a rent or not. Then the analyst uses those variables into their models in order to calculate the probability of default. And let's say that the corresponding probability equals 0.3. The next step is to compare this number with the predefined threshold of acceptance.

For example, a risk diverse institution would may not accept any loans if the score of the model would be greater than 0.15. Or a more risk seeking institution, would accept an application – maybe with a greater interest rate- if the corresponding score would be between 0.2 and 0.4. So, the loan is accepted or rejected according to the predefined threshold. In many cases, institution would probably accept bigger scores in order to avoid cases of false rejections; loans that were characterized as bad, but in the long run, the applicant would succeed to be on time on their obligations. Therefore, it clearly seems to be not only a matter of prediction modeling, but also the institution strategy and risk tolerance which drives decision making.

Chapter 4

A review of bankruptcy prediction models

Before deep-diving into the aspects of the different classification models used in practice for the calculation of credit score, we should first make a brief introduction of the existing research and implementations of different techniques during the last years in order to better understand the existing framework of credit scoring and machine learning applications in bankruptcy prediction. We have already introduced the historical evolution of credit scoring during the last fifty years; thus, the author of this Thesis believes that a proper review of bankruptcy prediction models would add incremental value on getting more insights into the subject of this Master Thesis.

Prediction of corporate failure using past financial data is a well-documented topic. One of the first researchers to study bankruptcy prediction was Beaver, back in 1966. (Beaver, 1966). He investigated the predictability of the 14 financial ratios using 158 samples consisted of failed and non-failed firms. The data Beaver used where extremely few when compared to our days' data availability, but back in 1966, Beaver was an innovator on his field and can be regarded as one of the first researchers on the field of predictive analytics. Beaver's study was followed by Altman's models based on Discriminant Analysis to identify the companies into known categories (Altman, 1968). Altman proposed that bankruptcy could be explained quite completely by using a combination of five (selected from an original list of 22) financial ratios. He utilized a paired sample design, which incorporated 33 pairs of manufacturing companies. Criteria for pairing those companies were based on size and industrial sector of activity.

The classification of Altman's model based on the value obtained for the Z score had a predictive power of 96% for predicting bankruptcy one year prior to the actual occurrence. However, the problem of those conventional methods was that they had specific restrictive assumptions such as the linearity, normality and independence among predictor or input variables. Taking into consideration that in the vast majority of financial data sets the violation of those assumptions for independent variables frequently occurred, Altman’s methods had limitations to obtain the effectiveness and validity.

Recently, a number of studies have demonstrated that artificial intelligence approaches that are less vulnerable to the aforementioned assumptions, can be alternative methodologies for classification problems to which traditional statistical methods have long been applied. While traditional statistical methods assume certain data distributions and focus on optimizing the likelihood of correct classification (Liang, Chandler, & Han, 1990), machine learning techniques – which will be thoroughly discussed in the next chapters - is a technology that automatically extracts knowledge from training samples by analyzing heavy volumes of data, in order to generate patterns and identify discrimination between sample characteristics. Therefore, the difference between a statistical approach and a Machine learning approach is that different assumptions and algorithms are used to generate knowledge structures.

One of the first complete research introduced in Machine Learning applications in credit scoring had been applied by Messier and William in 1998 (Messier & William, 1988). They extracted bankruptcy rules using rule induction algorithms that classified objects into specific groups based on observed characteristics ratios. They drew their data from two prior studies and began with 18 ratios. Their algorithm developed a bankruptcy prediction rule that employed five of these ratios. This method was able to correctly classify 87.5% of the holdout data set.

Shaw and Gentry (1990), applied inductive learning methods to risk classification applications and found that inductive learning’s classification performance was better than probit or logit analysis. They have concluded that this result can be attributed to the fact that inductive learning is free from parametric and structural assumptions that underlie statistical methods. Chung and Tam (1992), compared the performance of two inductive learning algorithms (ID3 and AQ) and Neural Networks (NNs) using two measures; the predictive accuracy and the representation capability. Results generated by the ID3 and AQ are more explainable yet they have less predictive accuracy than NNs. The predictive accuracy of ID3 and AQ is 79.5% while that of NN is 85.3%.

Neural Networks are capable of identifying and representing non-linear relationships in the data set, and thus, they have been studied extensively in the fields of financial problems including bankruptcy prediction. Those Machine Learning algorithms fundamentally differ from parametric statistical models. Parametric statistical models require the developer to specify the nature of the functional relationship such as linear or logistic between the dependent and independent variables. Once an assumption is made about the functional form, optimization techniques are used to determine a set of parameters that minimizes the measure of error (Seen & Lee, 2005). In contrast, NNs with at least one hidden layer, use data to develop an internal representation of the relationship between variables so that a priori assumptions about underlying parameter distributions are not required. As a consequence, better results might be expected with NNs when the relationship between the variables does not fit the assumed model (Shin, Taik, & Kim, 2005).

The first attempt to use NNs for bankruptcy prediction is found in Odom and Sharda's work in (1990). The model they had proposed had five input variables, the same as the five financial ratios used in Altman's study, and one hidden layer with five nodes and one node for the output layer. They took a research sample of 65 bankrupt firms between 1975 and 1982, and 64 non-bankrupt firms, overall 129 firms. Among those, 74 firms (38 bankrupt and 36 non-bankrupt firms) were used to form the training set, while the remaining 55 firms (27 bankrupt and 28 non-bankrupt firms) were used to make a test sample. Multivariate Discriminant Analysis was conducted on the same training set as a benchmark to compare the results of Neural Networks with an already successfully applied statistical method. As a result, NNs correctly classified 81.81% of the hold out sample while Multivariate Discriminant Analysis only achieved 74.28%.

Fletcher and Goss (1993), compared a NNs performance with a logit regression model. Their data were drawn from an earlier study and were limited to 36 bankrupt and non-bankrupt firms. Their model used three financial variables, and because of the very small sample size, they used a variation of the 18-fold cross-validation analysis. Although the NN models had higher prediction rates than the logit regression model for almost all risk index cutoff values, due to a very small sample size, the training effort for building NNs was much higher.

Zhang et al. (1999) also compared a NN models' performance with a logit model, and employed a five-fold cross-validation procedure, on a sample of manufacturing firms. The NNs significantly outperformed the logit regression model with accuracy of 80.46 versus 78.18% for small test set, and with accuracy of 86.64 versus 78.65% for large test set. Since the robustness and performance of the NN model improved significantly from small sets to large sets, user of NN would be well advised to use a large number of sets.

Support Vector Machines have also been applied in bankruptcy prediction modeling. SVM produces a binary classifier, the so-called optimal separating hyperplanes, through extremely non-linear mapping the input vectors into the high-dimensional feature space (Shin, Taik, & Kim, 2005). SVM constructs linear model to estimate the decision function using non-linear class boundaries based on support vectors (Hui & Sun, 2006). SVM trains linear machines for an optimal hyperplane that separates the data without error and into the maximum distance between the hyperplane and the closest training points.

Haardle and Schaafer (2003), compared Support Vector Machines with NNs and Multivariate Discriminant Analysis and resulted that SVM obtained the best results (70.35–70.90%) accuracy depending on the number of inputs used, followed by NN (66.11–68.33%) with MDA came last with only 60 –63.5% accuracy on predicting the probability of bankruptcy. The major disadvantage of Neural Networks and Support Vector Machines, is that their results are not easy to be understood business wise in terms of the process followed to extract the corresponding probabilities. Analysts refer to NNs and SVM as “black boxes” in terms of their inability to interpret their decision-making process. More about Support Vector Machine will be introduced in the next chapters of this Master Thesis.

Chapter 5

Machine Learning: A major breakthrough

5.1 Machine Learning: Definition and Concepts

Learning, like intelligence, covers such a broad range of processes that it is difficult to define precisely (NilSson, 1998). Almost thirty-five years ago, Michalski (Michalski, Carbonell, & Mitchell, 1983), in their book on Machine Learning and AI, pinpointed the different aspects and dimensions of learning. According to the authors, learning is a many-faceted phenomenon that includes the acquisition of new declarative knowledge, the development of motor and cognitive skills through instruction and practice and the discovery of new facts and theories through observation and experimentation. The last phrase is a very good initial to introduce the concept of Machine Learning.

Machine learning is a school of computer science that focuses on programming machines to improve their own performance through data and iteration. In other words, “through observation and experimentation”, a machine is trained to learn new things, to better understand already known concepts and to recognize patterns in a similar but more robust and effective way than any human beings. According to an extensive report of the Royal Society (2017), Machine Learning is the technology that allows systems to learn directly from examples, data, and experience. If the broad field of artificial intelligence (AI) is the science of making machines smart, then machine learning is a technology that allows computers to perform specific tasks intelligently, by learning from examples. These systems can therefore carry out complex processes by learning from data, rather than following pre-programmed rules.

Nilson (1998) argues that a machine learns whenever it changes its structure, program, or data (based on its inputs or in response to external information) in such a manner that its expected future performance improves. Some of these changes, such as the addition of a record to a data base, fall comfortably within the province of other disciplines and are not necessarily better understood for being called learning. But, when the performance of a speech-recognition machine for example improves after hearing several samples of a person’s speech, then, yes, we can definitely say that a machine can actual learn by examples.

Recent years have seen much discussion of machine intelligence, and what this means for our health, productivity, and wellbeing (Royal Society, 2017). Machine learning apparently promises to save lives, address global challenges and add trillions of dollars to the global economy through increasing productivity; while doing so it also fundamentally changes the nature of work, and shapes, or defines, the choices people make in everyday life. Between these extremes, there lies a potentially transformative technology, which brings with it both opportunities and challenges, and whose risks and benefits need to be navigated as its use becomes more central to everyday activities. As a result of these advances, systems which only a few years ago struggled to achieve accurate results can now outperform humans at specific tasks. There now exist voice and object recognition systems that can perform better than humans at certain tasks, though these benchmark tasks are constrained in nature. Many people now interact with machine learning-driven systems on a daily basis: in image recognition systems, such as those used to tag photos on social media; in voice recognition systems, such as those used by virtual personal assistants; and in recommender systems, such as those used by online retailers. In addition to these current applications, the field also holds significant future potential; further applications of machine learning are already in development in a diverse range of fields, including healthcare, education, transport, and more. Machine learning could provide more accurate health diagnostics or personalized treatments, tailor classroom activities to enhance student learning, and support intelligent transport systems. It could also support scientific advances, by drawing insights from large datasets, and drive operational efficiencies across a range of industry sectors

5.2 Machine Learning Historical review and major breakthroughs

Someone may think that the concept of Machine Learning is quite new as a field of research, let alone the implementation of ML in everyday problems. But the truth is totally different. The ideas behind machine learning have a long history, and rely on mathematics from hundreds of years ago and the enormous developments in computing in the last 70 years. In this chapter of this Master Thesis, we will go through a systematic historical review of Machine Learning algorithms' evolution through years, starting from the first recognized AI machine, reaching our days, where the notion of ML has been deeply penetrated into our lives. We will use an extensive collection of BBC articles available in BBC's free webpage library (BBC, BBC) in order to introduce the major breakthroughs of AI during the last 150 years.

Many of the mathematical underpinnings of modern machine learning predate computers and come from statistics. According to Hogenboom (2016), one of the preliminary breakthroughs in the field was the work of the great mathematician Thomas Bayes in the 18th century, which led Pierre-Simon Laplace to define Bayes' Theorem in 1812 (Hogenboom, 2016). Another great mathematician with recognized research in many fields of mathematics and statistics was Adrien-Marie Legendre. In 1805, Legendre developed the Least Squares method for data fitting. A technique that is commonly used in statistics with numerous applications such as the calculation of the best fit line in linear regression. Andrey Markov, in 1913, described analysis techniques later called Markov Chains. Markov Chains are broadly used nowadays and are considered as fundamental techniques to modern machine learning (Hogenboom, 2016).

In the late 1940s, work proceeded to develop stored-program computers that hold their instructions in the same memory used for data. Those computers may be considered as the first attempts to develop a machine that works in a similar way as the human brain in terms of how it processes the input data. The first computers of this category began the modern computing revolution (Hogenboom, 2016). The most well-known computers of this type were the Manchester Small-Scale Experimental Machine – commonly known as "Baby", the Manchester Mark 1 in 1949, and the University of Pennsylvania's EDVAC in 1951 (Norman, 2017).

Later, in 1950, the father of Artificial Intelligence, Alan Turing, published his work on *Computing Machinery and Intelligence*, in which he asked: *“Can machines think?”* According to Dr Andrew Hodges (2016), Turing’s growing understanding of the power of computers led to the publication of his paper which may be considered as *“one of the first attempts to describe how ‘artificial’ intelligence could be developed”* (Hodges, 2016). Turing is famous among the greatest innovators of all times for proposing the *“imitation game”*; a test to determine whether a computer was intelligent by asking a person to distinguish between a human and a computer when communicating with them both through typed messages (Hodges, 2016).

In another article of BBC, the work of Marvin Minsky and Dean Edmonds in the field of AI is explicitly described (BBC, 2016). Minsky and Edmonds, in 1958 have introduced the first artificial neural network – a computer-based simulation of the way organic brains work; It was the result of their research in Massachusetts Institute of Technology's Artificial Intelligence Lab in the late 50’s.

Minsky and Edmonds computer – known as *‘The Stochastic Neural Analog Reinforcement Computer (SNARC)*, learned from experience and was used to search a maze, like a rat in a psychology experiment. It was built along connectionist principles, representing a mind as a network of simple units within which intelligence may emerge (BBC, 2016). Minsky went on to work at the MIT Artificial Intelligence Laboratory and made many other significant interventions in the AI debate. He was an advisor on the film 2001: A Space Odyssey.

Later in 1960, the Backpropagation technique was introduced, consisting another major benchmark and significant breakthrough in the field of ML. First described in the 1960s as part of control theory and adopted for neural networks, backpropagation fell out of favor until work by Geoff Hinton and others using fast modern processors demonstrated its effectiveness. Deep learning nets are now a mainstay of current machine learning. In 2017 Hinton, who now works for Google, expressed concerns that backpropagation has reached its limits in building machine learning systems and that new insights from biology are needed.

Public awareness of AI increased greatly when the power of Artificial Intelligence entered the world of chess. IBM computer named Deep Blue beat world chess champion Garry Kasparov in the first game of a match (Straseer, 2017).

Kasparov played against the computer in 1996 and managed to win the game, but in 1997 a brand new and upgraded version of Deep Blue won the famous world champion indicating the start of a whole new age in the field of computing power. Deep Blue “played” an impressive game of chess, largely relied on brute computing power to achieve this, including 480 special purpose ‘chess chips’. The machine had the ability to react better than Kasparov by searching from 6-20 moves ahead at each position, having learned by evaluating thousands of old chess games to determine the path to win the match.

If Deep Blue’s chess expertise was the big AI story of the last millennium, then AlphaGo’s success at Go has replaced it in popular culture.

Developed by DeepMind researchers, AlphaGo won its first match against a professional in 2015, beat the world’s number two player Lee Sedol in March 2016 and the number one player Ke Jie in 2017. AlphaGo’s neural network is trained by playing both humans and computers, and uses a Monte Carlo tree search algorithm to find moves. Its success is significant as AI researchers consider the game of Go to be a hard problem and had not anticipated that humans would start losing to computers so soon.

By reviewing the history and evolution of ML through the aforementioned timeline of major breakthroughs, the main conclusion point is that the power of Machine Learning algorithms rely on their ability to train themselves on real data and learn from consecutive repetitions and trials, searching patterns and returning findings that a human brain is unable to demonstrate due to certain constrains of computing power. Therefore, the next think which comes under consideration is how a machine learns and how training offers incremental predictive power, over performing almost every human calculation technique. In the next chapter, we will describe the notion of “training”, indicating the reasons why a machine should train itself in order to gain its predictive power.

5.3 Learn through training

One might ask “Why should machines have to learn? Why not design machines to perform as desired in the first place?” There are several reasons why machine learning is important. Nilson (NilSson, 1998) provides as with a few reasons why it is necessary for a machine to learn by example and why it is almost impossible to construct a machine that knows everything a priori:

- Some tasks cannot be defined well except by example; that is, we might be able to specify input/output pairs but not a concise relationship between inputs and desired outputs. We would like machines to be able to adjust their internal structure to produce correct outputs for a large number of sample inputs and thus suitably constrain their input/output function to approximate the relationship implicit in the examples.
- It is possible that hidden among large piles of data are important relationships and correlations. Machine learning methods can often be used to extract these relationships (data mining).
- Human designers often produce machines that do not work as well as desired in the environments in which they are used. In fact, certain characteristics of the working environment might not be completely known at design time. Machine learning methods can be used for on-the-job improvement of existing machine designs.
- The amount of knowledge available about certain tasks might be too large for explicit encoding by humans. Machines that learn this knowledge gradually might be able to capture more of it than humans would want to write down.
- Environments change over time. Machines that can adapt to a changing environment would reduce the need for constant redesign.
- New knowledge about tasks is constantly being discovered by humans. Vocabulary changes. There is a constant stream of new events in the world. Continuing redesign of AI systems to conform to new knowledge is impractical, but machine learning methods might be able to track much of it.

So, it is more than clear that the concept of machine learning can be summarizes in the following sentence; Provide the machine with tones of appropriate data and let it learn by examining patterns and correlations between variables; patterns that a human is unable to recognize due to lack of computing power.

5.4 Types of Machine Learning Algorithms

Machine learning algorithms, also known as statistical learning algorithms, perform tasks without being explicitly programmed (Damrongsakmethee & Neagoe, 2017). These algorithms can be separated in two forms, supervised and unsupervised machine learning. In **(Figure 4),** we can see the different types of machine learning algorithms and the corresponding problems these algorithms best fit to:

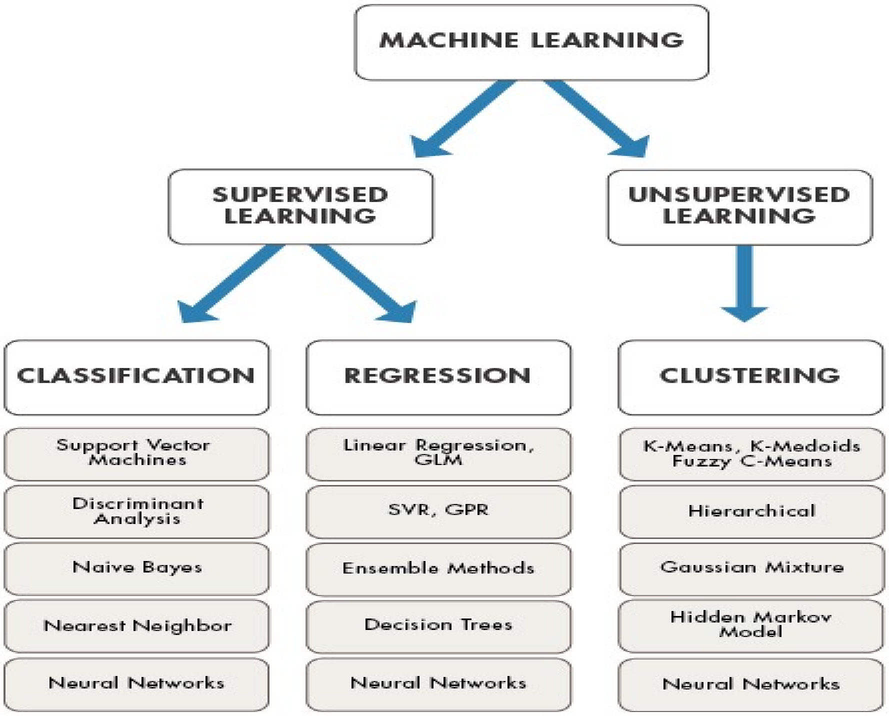


Figure 4: Types of Machine Learning Algorithms

Supervised machine learning algorithms learn to predict an outcome, the response, based on the values of different features or variables (NilSson, 1998). The data used to create a model with a supervised learning algorithm is historical and thus contains known response values. Supervised learning is used in a variety of applications, such as speech recognition, spam detection and object recognition. The goal is to predict the value of one or more output variables given the value of a vector of input variables x (Musumeci & Rottondi, 2018). The output variable can be a continuous variable (regression problem) or a discrete variable (classification problem). A training data set comprises N samples of the input variables and the corresponding output values. Different learning methods construct a function that allows to predict the value of the output variables in correspondence to a new value of the inputs.

In contrast, unsupervised machine learning algorithms are models applied to data where there is no response value known. The performance of these models is hard to evaluate, because there are no observations to test predictions on (NilSson, 1998). Unsupervised machine learning is used to learn relationships and find structure in data. The problem in this case, typically, is to partition the training set into subsets, in the most appropriate way. Unsupervised learning methods have application in taxonomic problems (clustering -segmentation) in which it is desired to invent ways to classify data into meaningful categories (NilSson, 1998).

5.5 Machine Learning and Credit Risk Assessment

Now that we have introduced the concept of Machine Learning and described the key characteristics of Machine Learning algorithms, we will focus on the use of AI in credit scoring trying to better understand the incremental value of Artificial Intelligence towards the prediction of the probability of default; the final challenge for every credit analyst.

There have been major advances in the application of Machine Learning in the recent past due to a plethora of industry drivers that have revolutionized the utilization of these techniques in the risk management sphere, and beyond. Moody and Haydon (Moody & Haydon, 2018) on their article in credit risk and Machine Learning are analyzing four reason that clearly demonstrate the swift towards ML in dealing with large volumes of data. According to the authors, the first reason is the large expansion of data during the last decade in several dimensions; size, velocity and variety. Simultaneously the abilities to record, store, combine and then process large datasets from many disparate sources has experienced wholesale improvements. This is not limited to just traditional sources, but also alternative data which fueled the need to extract information value from these sources. Machine Learning models by nature are able to deal with big data in very efficient ways and thus, ML algorithms should be regarded as excellent tools for using in complex prediction problems, as, for example, credit scoring.

Second, the ease of access to enhanced computational efficiency through hardware that can run specialized operations in large scale, and also in coding language enhancements which have moved towards functional programming, *"have transformed the game in terms of integrating Machine Learning techniques."* (Moody & Haydon, 2018).

Third, reproducible research and analysis has been widely adopted by the data science community. This is defined as a set of principles about how to do quantitative and data science driven analysis, where the data and code that leads to a decision or conclusion should be able to be replicated in an efficient and clear way (Moody & Haydon, 2018).

Finally, the pervasiveness of Open Source libraries, packages and toolkits has opened doors for the community to contribute via teams of specialists, sharing code base and packaging them into easy and modular functions.

The evolution of statistical languages and toolkits should be regarded as another benchmark towards the excessive use of Artificial Intelligence in predictive modelling. Languages for example, such as R, and Python, have become excellent hubs for numerical computing providing the analyst with a majority of different methods to apply in order to easily built a ML algorithm. Additionally, the huge number of ready to use libraries and the extensive documentation available in the Internet, are forming a fertile ground for an individual to

All those improvements to machine learning in recent years have led to many financial institutions leveraging machine learning to produce not only higher returns but also less risk on investments. This includes firms implementing high-frequency trading desks that leverage machine learning to help make trading decisions in a fraction of the time it would take a person. Alongside this, many more user-oriented financial tools have become available with the help of machine learning. These include services such as fraud detection technologies, and insurance underwriting All of these services would be much more difficult, if not impossible without machine learning.

Mackenzie (Mackenzie, 2018), argues in favor of using Machine Learning for risk evaluation. He mentions that *“leveraging machine learning for various uses including risk management can only stand to benefit financial institutions”*. He also argues that not only financial institution like banks and assurance companies, but also *“we, the borrowers”* benefit from lower rates on loans and an improved sense of security that our bank will stay solvent (Mackenzie, 2018).

Credit risk assessment and modeling is one of the most important topics in the field of financial risk management (Wang, Wang, & Lai, 2005). Due to recent financial crises, credit risk assessment has been the major focus of financial and banking industry.

Mackenzie (2018) argues that having machine learning in financial institutions will also reduce the chance of future economic disasters caused by lapses in human judgement such as the *“great depression”* in 2008. Ever since the financial crisis, banks around the world have put a more significant emphasis on risk management systems in the effort to reduce the chances of another global economic recession (Backman & Zhao, 2017). From these systems, credit risk was one of the main sub-causes of the 2008 financial crisis which became of great importance.

Especially for any credit-granting institution, such as commercial banks and certain retailers, the ability to discriminate good customers from bad ones is crucial (Wang, Wang, & Lai, 2005). To a bank, whether a client can deal with their obligations, is the difference between making a profit from interest and otherwise having to liquidate the client’s assets or even lose the loan entirely. Mackenzie (2018) concludes that, finding ways to analyze creditworthiness and making smart risk management decisions is a top priority for the banks. By utilizing big data and machine learning the financial institutions can calculate the risks on loans and other financial transactions to a much greater degree, which in turn will help alleviate the overall risks the banks take with customer’s money.

Although applications of machine learning and big data will never be able to completely irradiate the potential risks that stem from lending to risking clients, it is a significant step in the right direction that will help reduce future financial losses and possible disasters.

The need for reliable models that predict defaults accurately is imperative so that the interested parties can take either preventive or corrective. Additionally, with the rapid growth in the credit industry, credit scoring models have been extensively used for the credit admission evaluation

(Thomas L. C., 2000). During the last decades, several quantitative methods have been developed for the credit admission decision. The credit scoring models are developed to categorize applicants as either accepted or rejected with respect to the applicants' characteristics such as age, income, and marital condition.

Credit officers are faced with the problem of trying to increase credit volume without excessively increasing their exposure to default. Therefore, to screen credit applications, new techniques should be developed to help predict credits more accurately. Machine Learning techniques and algorithms are systematically gain more ground in the informal race between credit risk evaluation methods due to their ability to over perform – in most cases- the majority of rule-based decision-making processes commonly used in the financial sectors during the last decades.

Managers have stopped trying to learn from experience and have relied their decision making into much more reliable and objective techniques searching for answers that would minimize their long-term exposure to risk. An unbiased objective prediction of a borrower's probability of going bankrupt can be a useful management tool.

The major advantage of ML algorithms is their ability to process large volumes of data regardless the source or even the implicit correlations between different aspects and characteristics that a human eye – regardless the level of expertise – is capable to recognize. While not approaching the human-level intelligence which is usually associated with the term AI, the ability to learn from data increases the number and complexity of functions that machine learning systems can undertake, in comparison to traditional programming methods (Royal Society, 2017).

Machine learning can carry out tasks of such complexity that the desired outputs could not be specified in programs based on step-by-step processes created by humans. The learning element also creates systems which can be adaptive, and continue to improve the accuracy of their results after they have been deployed. The latest phrase is the key to comprehend the incremental value of Machine

Learning techniques in the field of credit risk calculation. As new data become available in a daily bases and characterization of what may consist a significant variable for the estimation of credit worthiness change, models that can adapt rapidly on those changes should be regarded as extremely useful to deploy and rely on in order to better understand the implicit risks and adjust the decision making accordingly.

Numerous methods have been proposed for bankruptcy prediction. Some review papers have attempted to categorize them into statistical methods, intelligent systems, data mining, and machine learning techniques. However, the boundaries between these disciplines are slowly vanishing (Murphy, 2018); statistical methods like logistic regression and intelligent systems such as support vector machines are now taught in almost every machine learning course. Therefore, all these data-driven learning methods for continuous and discrete outputs will simply be considered as machine learning techniques

Chapter 6

Credit scoring statistical techniques

As we have already discussed in the previous chapters, the main idea of credit scoring modelling is to identify the features that influence the payment or the non-payment behavior of the costumer as well as his default risk, occurring the classification into two distinct groups characterized by the decision on the acceptance or rejection of the credit application. (Louzada, Anderson, & Guilherme, 2016). A wide range of statistical techniques are used in building the scoring models. Most of these statistical, and some of these non-linear, models are applicable to build an efficient and effective credit scoring system that can be effectively used for predictive purposes (Thomas L. C., 2000). In this chapter, we will go through a detailed review of the literature on statistical techniques used in order to calculate credit scoring.

6.1 Classification algorithms

Classification is one of the Data Mining techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class such that classification error will be least. It is used to extract models that accurately define important data classes within the given dataset (Nikam, 2015). Credit scoring shall be treated as a classification problem with two mutually exclusive classes; the borrower will default or not. In most cases, those two events are being assigned with a probability, which defines the class that the person belongs. Therefore, the problem of assigning a proper credit score to an individual is an issue of classifying those individuals into one of the two aforementioned groups. The better the results of the classification method, the greater the probability to predict the “behavior” of the individual in terms of default.

6.2 Statistical methods used in practice

Overall, the main classification methods in credit scoring and those that are introduced in this master thesis are neural networks (NN) (Ripley, 1996), support vector machine (SVM) (Vapnik, 1998), linear regression (LR) (Hand & Kelly, 2002), decision trees (Breiman, 1996), logistic regression (LG) (Abdou, Pointon, & El-Masry, 2008), (Sohn, Dong, & Yoon, 2016), discriminant analysis (Fisher, 1986) and K Nearest Neighbors (Mukid & Widiharih, 2018), (Hand & Henley, 1997). In what follows, the author of this Master Thesis provides a brief review of those different classification algorithms in terms of the basic functional components of each different method.

6.2.1 Discriminant Analysis

Discriminant analysis and linear regression have been the most widely used statistical techniques to building scoring cards (Hand & Henley, 1997). Discriminant analysis is a statistical technique which allows the researcher to study the differences between two or more groups of objects with respect to several variables simultaneously. Fisher introduced the concept of discriminant analysis in his work on iris classification problem (Fisher, 1986). Discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature (Unknown Author, 2018). The term categorical variable means that the dependent variable is divided into a number of categories. For example, three brands of computers, Computer A, Computer B and Computer C can be the categorical dependent variable. In the case of credit scoring, the categorical variable is whether the borrower will default or not. When a new customer is applying for a loan, the bank must decide whether to grant him or not the requested loan by applying a discrimination rule. As a result of this process, the applicant will receive a score which classifies the application in one of the existing categories (e.g. bad payers, good payers). The discrimination rule offers support for decision of granting or not granting a loan, by attending at the background of the applicant and providing the required risk assessment (Mircea & Pirtea , 2011).

6.2.2 Linear regression

Linear regression methods have become an essential component of any data analysis concerned with describing the relationship between a response (dependent) variable and one or more independent variables. According to Yan (2009), linear regression models was the first type of regression analysis to be studied rigorously, and to be used extensively in practice. The reason is that models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine (Yan, 2009). Linear regression is quite easy to understand as it does exactly what its description says; the target value is estimated through a linear combination of all explanatory variables given that variables are linear related with each other. For example, if we have 3 explanatory variables X_1, X_2, X_3 and our target value is denoted as Y, then the linear regression algorithm will return a linear equation that describes Y, as a linear combination of X_1, X_2, X_3 with their corresponding coefficients b_1, b_2, b_3 and b_0, which is the point of interception of the regression line with the y-axis:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

A set of values from the vector of explanatory variables are provided in the linear regression equation, which returns an estimation of the target variable for the given values of the vector.

Coefficients are calculated through a process which is known as *"least squares approach"*. This approach takes into consideration the minimization of error between the real and the predicted value of the target variable. In other words, the best fit line is the one that minimizes the sum of square errors between the distances of the real and the predicted values through the regression model. The less the error, the better the regression line **(Figure 5).**

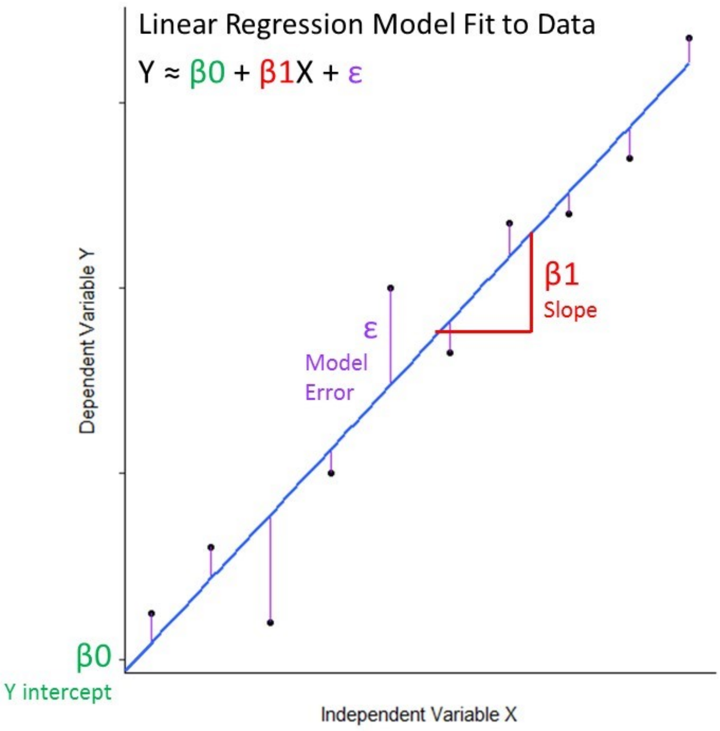


Figure 5: The best fit line is the one that minimizes the errors between real and predicted values (ε)

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression can be applied to binary classification problems such as credit scoring by defining the actual values of the dependent variables as categorical variables. For example, linear regression might be used in credit scoring to express the probability that an applicant for credit will not default based on a set of variables or features associated with applicants for credit. In this application, pi may represent the predicted probability that applicant i, i=1,2,….,m, will not default, with the actual value for the probability of no default be 1 if an applicant has not defaulted (i.e. a “good” applicant) and 0 if an applicant has defaulted (i.e. a “bad” applicant). Factors, such as customers’ historical payments, guarantees, default rates in a timely manner, can be analyzed by credit analysts, with linear regression to set up a score for each factor, and then to compare it with the bank’s cut-off score. If a new customer’s score passes the bank’s score, the credit will be granted (Thomas L. C., 2000).

An obvious weakness in using linear regression in binary classification problems is that it can produce predicted probabilities that are greater than 1 or less than zero. Pampel (2000), points that a major drawback of linear regression is that it is based on the assumption that the dependent variable and the residuals are following the normal distribution. However, in most cases in a binary classification problem, variable cannot be distributed normally as there are only two values for the dependent variable. According to Orgler (1970), linear regression produces models similar to those produced by discriminant analysis in binary classification problems.

Thomas (2000), describes the use of linear regression in the construction of scorecards, mainly because of its simplicity and the widespread availability of appropriate software. Orgler (1970) used linear regression to develop a model for evaluating commercial loans. However, linear regression will not be used in the model comparisons in this thesis because of its underlying assumptions.

6.2.3 Logistic Regression

Logistic regression, like discriminant analysis, is also one of the most widely used statistical techniques in the field. Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits **(Figure 6).**

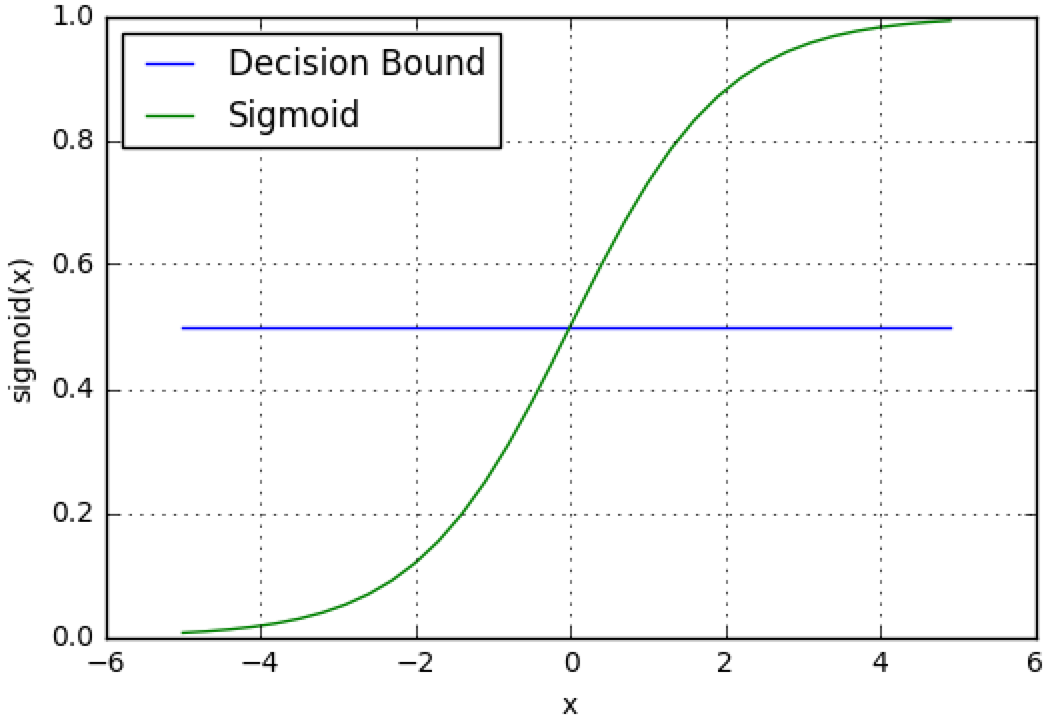


Figure 6: The Sigmoid function applied in logistic regression

The shape of the logistic function differentiates it from any linear equation making this function a good choice for modelling outcomes in a binary – not continuous – universe. Linear equation are not suitable predictors for binary variables as for values outside of the corresponding original regression domain, predictions will be outside of the 0-1 constraint. Logistic regression solves this problem by converging to either 0 or 1 any values that are outside the original regression domain. Another reason is that binary data don't satisfy the constant variance assumption of linear regression **(Figure 7):**

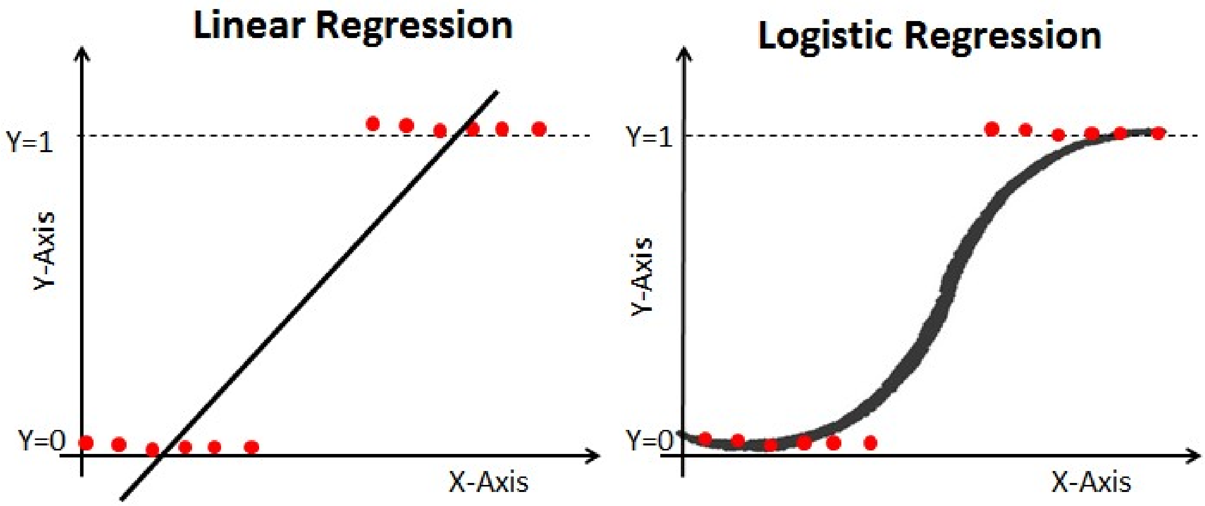


Figure 7:Sigmoid function converges outbound values

What distinguishes a logistic regression model from a linear regression model is that the outcome variable in logistic regression is dichotomous (a 0/1 or Y/N outcome). This difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions made when training the model. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression (Abdou & Pointon, 2011). In order to understand how logistic regression works, let us consider the binary classification problem with two classes, denoted 0 and 1, and assume there are m observations of known class membership, in other words, we have m observations already classified in one of the two binary categories. For observation $i, i=1,......,m$, let y_i with value 0 or 1, denote its class membership and let p_i denote the corresponding predicted probability of membership of class 1, so that $\frac{p_i}{1-p_i}$ represents the predicted probability of membership class of 1. The logistic regression model is then summarized in the following linear equation:

$$ln\left[\frac{p_i}{(1-pi)}\right] = b_0 + b_1X_{i1} + b_2X_{i2} + \ldots + b_nXn \quad i=1,\ldots\ldots,n\,.$$

So, the categorization of an observation into the appropriate class depicts on the calculation of the probability $\frac{p_i}{1-p_i}$, where:

$pi = exp\left(b_0 + b_1X_{1i} + \ldots + bnX_{1n}\right)/\left[1 + exp\left(b_0 + b_1X_{1i} + \ldots + b_nXn\right)\right]$, and coefficients bj, j=0, 1,…, n, are estimated using an iterative procedure to maximize the likelihood estimator. Therefore, every single observation within the population is attached with a score, a probability between zero and one. Unlike parametric methods like Linear Discriminant Analysis, logistic regression does not require assumptions about the population. Hand and Henley (Hand & Henley, 1997) indicate that on theoretical grounds, logistic regression may be a more proper statistical instrument than linear regression, given that the two classes, "good" loans and "bad" loans have been described. There is extensive literature on using logistic regression in building credit scoring applications. For example, Abdou, et al. (Abdou, Pointon, & El-Masry, 2008) used logistic regression and neural networks to investigate the ability of those models in evaluating credit risk in Egyptian banks applying credit scoring models. Crook et al, (Crook, Edelman, & Thomas, 2007) used a logistic regression model to evaluate the riskiness of lending to a credit applicant. More recently, Sohn et al. (Sohn, Dong, & Yoon, 2016), employed a logistic regression model in order to relate the probability of a loan default of the firms with several evaluation attributes associated with technology.

The main drawback of logistic regression is the model parameters must be estimated using an iterative maximum likelihood procedure that requires more computations than, for example, linear regression (Thomas L. C., 2000), although this problem has been reduced by improvements in computing technology. In addition, Thomas (2000) argues that as with linear regression, logistic regression is sensitive to correlated independent variables. One of the strengths of logistic regression is that, as with discriminant analysis and linear regression, it allows the user to identify the features that are good predictors of the dependent variable. It is therefore possible to produce a parsimonious model with the same (or better) performance as the model containing all the possible features. In general, logistic regression is a practical and easy-to-use method that can produce good results in building classification models.

6.2.4 Nearest Neighbors

Nearest neighbor methods, such as the k-nearest neighbor (k-NN) method, are nonparametric methods of estimating the probability of class membership from a set of values of features associated with an observation or object (Abdou & Pointon, 2011). The classification of an observation into the proper class is decided by the algorithm in a very simple, yet, explicit way. The probability, $P(Y/X)$ of membership of class Y for an observation or object of unknown class with vector of feature values X may be given by the proportion of its K nearest neighboring observations of known class membership that belong to class Y (Abdou & Pointon, 2011). In the k-NN method, the parameter k, which defines the size, but not the shape, of a neighborhood, and a separation metric for assessing proximity must be specified. In most cases, Euclidean distance is used as the separation metric in K-NN method but also, other more complex metrics in which different weights are attached to each dimension are also applied to define the most suitable grouping for each observation.

As we have already mentioned, K Nearest Neighbors is a nonparametric method, which means that it does not make any assumptions about the probability distribution of the input. According to Mukid (2018), the main idea of k-NN algorithm is that whenever there is a new point to predict, its k nearest neighbors are chosen from the training data. Then, the prediction of the new point can be the average of the values of its k nearest neighbors (Mukid & Widiharih, 2018).

In order to simplify how k-nearest neighbors' model works, let us consider that we have only two explanatory variables X_1, X_2 and we want to categorize every observation in a proper category. And let's say that we have 3 categories into which every observation should be grouped. The algorithm starts by calculating the distance of every observation from a pre-defined point in each group, called centroids (C_1, C_2, C_3) **(Figure 8):**

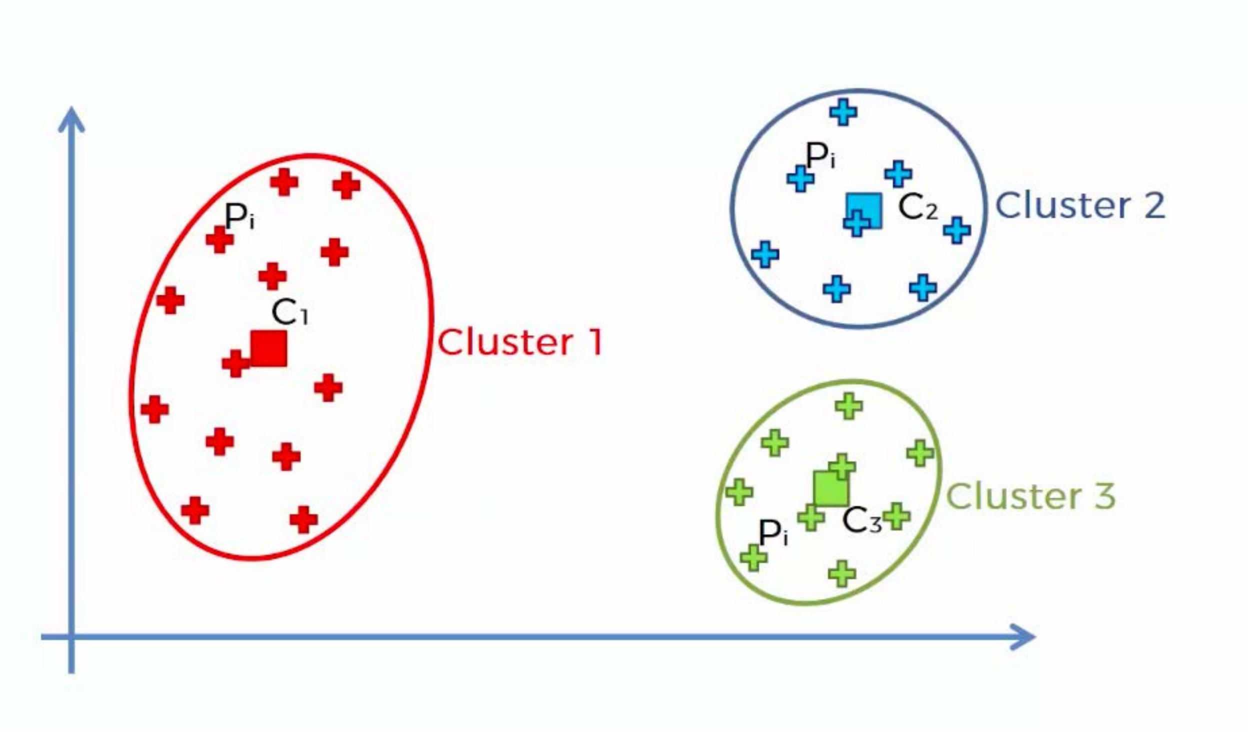


Figure 8: A typical clustering example using K Nearest Neighbors

The Euclidean distance of every observation from the 3 centroids is calculated and the observation is attached in the group in which the distance of the observation from the centroid is minimum. In the same way, every observation is grouped in one of the 3 categories. Therefore, all observations are initially grouped in the 3 given categories. The process is repeated but now, the centroids are recalculated into each group of observations. Then again, the distance of every observation from the new centroids is calculated and the observation either stays attached in its preliminary group or changes group if the distance from another's group centroid is smaller than the distance from the centroid of the group that the observation is attached into. The process continuous until no observation changes the group they belong to.

There is an extensive research and applications of the K- Nearest Neighbors approach in credit scoring calculation. Hand and Henley (1997), used K-NN to distinguish between good or bad risk applicants. According to the researchers, K-NN method is suitable for credit scoring and is easy to apply (Hand & Henley, 1997). For a new applicant for credit, let KG and KB, where KB=K–KG, denote the number of good and bad cases respectively in the K design-set cases of known good and bad status nearest to the new case, as determined by the separation metric. The estimates of the probabilities that a loan is good ($P(G/X)$) or bad ($P(B/X)$) are then given by KG/K and KB/K respectively and the new observations are classified into the proper class H where KH is the maximum between KG and KB.

K-NN can also be updated as the population of applicants' changes and it is fairly easy to incorporate misclassification costs (Hand & Kelly, 2002). Mukid and Widiharih (2018), used a K-NN weighted approach to calculate credit score for a sample of 948 applicants in an Indonesian Bank out of which 184 were characterized as bad. The bank defined that a bad customer is someone who had missed three consecutive months of payments. The data consisted of 8 continuous explanatory variables including age, working experience, total income, other loan, and net income, interaction to bank, savings, and debt ratio. Their weighted K-NN model succeeded over 85% of accuracy in terms of distinguishing the applicants between good and bad borrowers. Hand (1997), argues in favor of K-NN algorithm for each ability to provide reasons for refusing credit, which may be a legal requirement, as the neighbors can provide a case-based explanation. They also found that k-NN classifiers compared favorably with linear regression, logistic regression, and classification trees in credit scoring (2005).

6.2.5 Classification Trees

A classification tree, or recursive partitioning, is a nonparametric classification approach in which observations are split into sets of similar class membership using appropriate tests or splitting rules. Classification trees can be represented by a tree diagram, such as the binary tree, i.e. a tree in which there are two branches at each node other than the terminal nodes, as appears in **(Figure 9).** The non-terminal nodes, represented by circles, in a classification tree specify a test to split observations into different subsets and the branches at non-terminal nodes represent the outcomes associated with the test. The top node is the root of the tree and a class label is associated with each leaf or terminal node (denoted by a square). The splitting rules in a classification tree can be based on simple comparisons or metrics such as the Kolmogorov-Smirnov statistic (Thomas L. C., 2000). The classification and regression tree (CART) proposed by Breiman (1996) is an example of a classification tree.

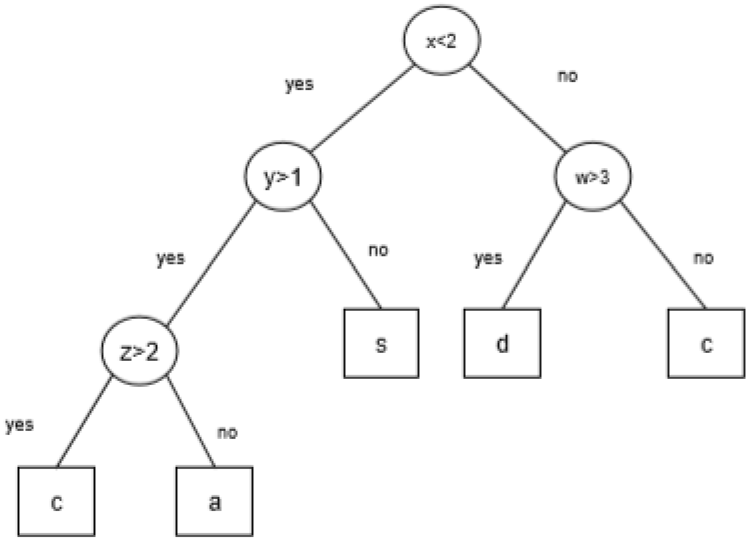


Figure 9: Example of classification tree. Squares denote possible outcomes and circles represent decision nodes

In using a binary classification tree for credit scoring, a design sample of applicants of known default risk is first split into two subsets, where each subset is composed of applicants with more similar default risk than the complete set of applicants. Each of these two subsets is then split into two using a different splitting rule to generate two more similar subsets in terms of default risk. This process of repeatedly splitting subsets of applicants into two is repeated until further subdivision does not yield more homogeneous subsets. In other words, when a terminal node is generated through the process (Abdou & Pointon, 2011). The tree can then be used to classify a new applicant, where for a new applicant with a specified vector of feature values, the predicted probability of low risk is given by the proportion of good applicants in the subset of the design sample at the terminal node associated with this vector of feature values. Abdou & Pointon enumerate three reason why classification trees are very suitable for use in credit scoring (Abdou & Pointon, 2011):

- First, the underlying decision process can be represented in a sequential way rather than simultaneously as is the case with other methods as for example, linear discriminant analysis or logistic regression.
- Second, the construction of nonlinear classifiers is very easy using a tree-based evolution process and
- Decision trees have the ability to handle both categorical and nominal variables.

Classification tree methods use historical data to construct so-called decision rules organized into tree-like architectures. In general, the purpose of this method is to determine a set of if-then logical conditions that permit prediction or classification of cases (Louzada, Anderson, & Guilherme, 2016). Bhatia et al. (Bhatia, Sharma, & Burman, 2017) pinpoints the usefulness of classification trees in credit scoring calculation. They state that tree-based learning algorithms like Decision Trees are considered to be one of the best and mostly used in the category of supervised learning methods. Tree based methods encourage predictive models with stability, high accuracy and easy of exploration. But how exactly do they work? In fact, tree-based methods map the non-linear relationships with a good accuracy. This method breaks down the dataset into smaller and smaller subsets of data while in the same period an associated decision tree is developed in an incremental manner. The system considers all possible splits to find the best one, and the (winning) sub-tree is selected based on its overall error rate or lowest cost of misclassification (Abdou & Pointon, 2011).

However, classification trees can become very large during their evolution until reaching on the terminal nodes where a proper classification of every observation to a given class is derived. Safavian and Landgrebe (1991), mention that in order to avoid such large and yet not easily interpreted classification trees, most approaches use a fixed design or training set and additionally, tree redesign may be required as additional data become available. An additional disadvantage of classification trees is that continuous variables are implicitly discretized by the splitting process, with information lost in this process (Louzada, Anderson, & Guilherme, 2016).

We have described in detail the statistical methods used in practice in classification problems, indicating their advantages and disadvantages and their implicit mechanisms for classifying observations into given datasets. In the next chapter, we will go through a similar review of the second major and broadly known category of classification algorithms, the Machine Learning approach, describing two of the most well-known and used ML algorithms; Neural Networks and Support Vector Machines.

6.3 Machine Learning Methods used in practice

Machine learning methods have been used in many classification tasks, e.g. Piramuthu (1999b), Shaw and Gentry (1988), Wang et al (2005). Methods such as neural networks, support vector machines and expert systems are less restrictive than many statistical methods as they do not require assumptions about the data used to build a model. However, these methods use a "black-box" approach for classifier construction and since information on the steps followed in deriving the weights for each feature is not produced, it is generally not possible to provide an interpretation of the results.

6.3.1 Neural Networks

Neural networks are mathematical techniques motivated by the operations of the human brain as influential in problem solving techniques (Abdou & Pointon, 2011). Gately (1996) was one of the first who provided a definition of neural networks referring to them as a problem connected with machine learning. He defined neural networks as *"an artificial intelligence problem solving computer program that learns through a training process of trial and error"*. Therefore, neural networks' building requires a training process, and the training procedure helps distinguish variables for a better decision-making outcome.

In an artificial neural network (ANN), each characteristic is taken as an 'input' and a linear combination of them is taken with arbitrary weights. The structure of a very simple multilayer network is shown in **(Figure 10).** The central column of circles is a hidden layer and the final circle is the output layer. The characteristics are linearly combined and subject to a non-linear transformation represented by the g and h functions, then fed as inputs into the next layer for similar manipulation (Crook, Edelman, & Thomas, 2007).

The final function yields values which can be compared with a cut-off for classification (Sohn, Dong, & Yoon, 2016). Each training case is submitted to the network, the final output compared with the observed value (0 or 1 in this case) and the difference, the error, is propagated back though the network and the weights modified at each layer according to the contribution each weight makes to the error value.

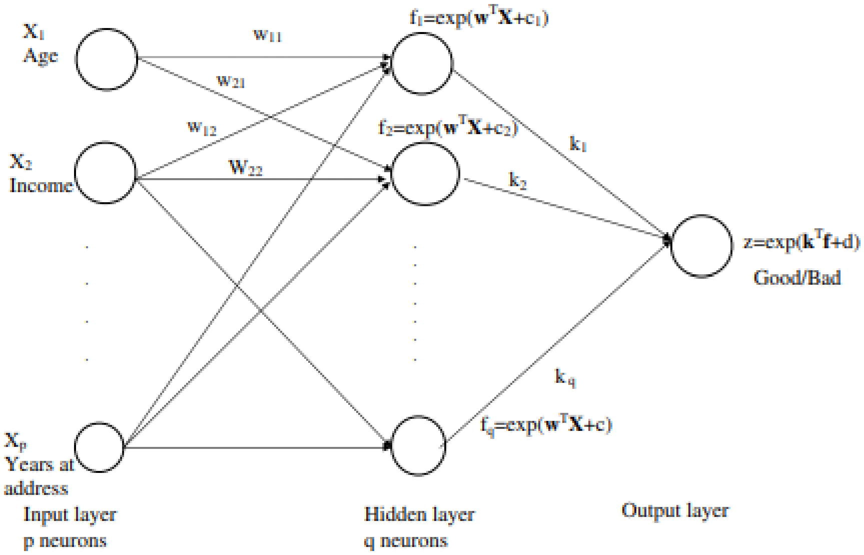


Figure 10: A two-layer neural network

The most significant difference between Artificial Neural Networks (ANN) and other classification algorithms like logistic regression and decision trees is that if the outcomes are unacceptable, the estimated scores will be changed by the nets until they become acceptable or until having each applicant's optimal score. The neural network learns by repeated adjustment of the weights. The difference between the output of the network and the target output can be seen as an error which we want to minimize (Landajo, de Andres, & Lorca, 2007). This can be done with the backpropagation algorithm, which starts at the output layer and propagates the error backward through the hidden layers and adjust the weights, usually by the use of some form of gradient descent where the weights are updated according to.

In other words, the networks are “intelligent” in auto correcting themselves by processing the outcome as feedback and using this outcome as input in a proper layer. Recently neural nets have emerged as a practical technology, with successful applications in many fields in financial institutions in general, and banks in particular.

6.3.2 Support Vector Machines (SVM)

Support vector machines (SVMs) were first proposed by Vapnik (1998) as learning systems for binary classification. SVM is a relatively new artificial intelligence method that is based on the structural risk minimization principle rather than the empirical risk minimization principle in order to determining the location of decision boundaries also known as hyperplane that produce the optimal separation of classes in a classification problem (Nikam, 2015). SVMs are trained using an algorithm from optimization theory and statistical learning theory to derive a separating hyperplane in a high dimensional feature space (Shawe-Taylor & Cristianini, 2000) .

The original model proposed by Vapnik (1998) was a linear classifier, but other types were later proposed in order to improve the accuracy of the original model. The main difference of the new models compared to the initial model is the function used to map the data into a higher dimensional space. New functions were proposed, namely: polynomial, radial basis function (RBF) and sigmoid. All these functions transform the original data into a higher dimensional space and then the linear classifier is used subsequently.

In **(Figure 11),** we can see an example of support vector machines. The green hyperplane has classified data into two categories, red and blue. In contrast with linear repressors, SVM’s are producing hyperplanes in a 3 or multidimensional space in order to properly classify each observation in the corresponding class.

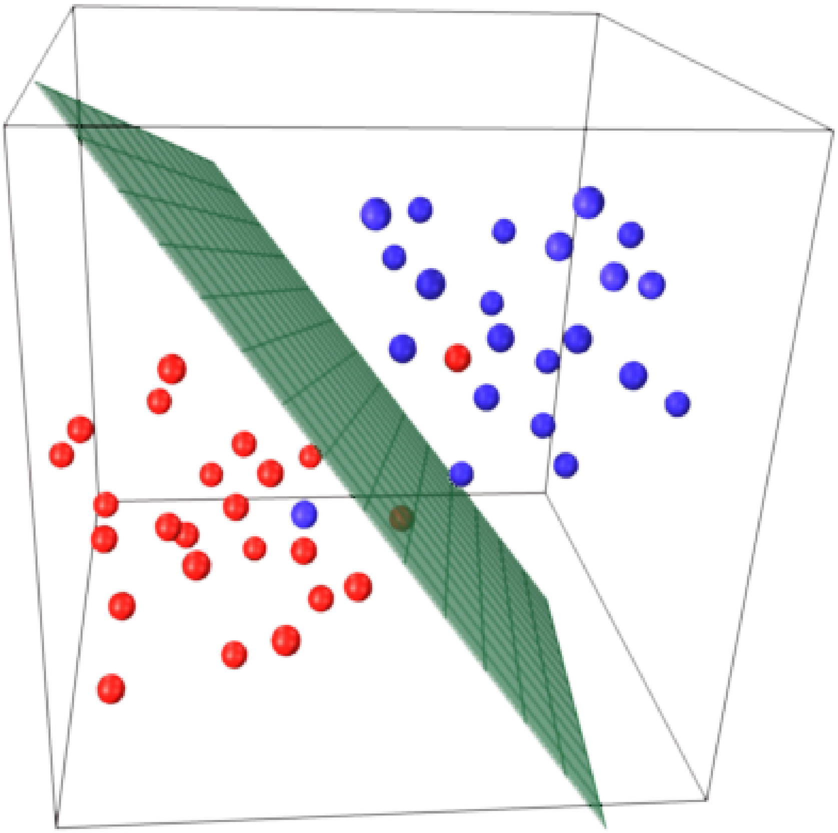


Figure 11: SVM classification: A Hyperplane is produce to classify observations into the corresponding class.

SVMs are based on a nonlinear mapping of the problem data into a higher dimension feature space (Shawe-Taylor & Cristianini, 2000). Wang et al. (Wang, Wang, & Lai, 2005) in their research, arguing towards the efficiency of SVM in classification problems. They state that SVM is a powerful and promising data classification and function estimation tool because of its ability not to run into over-fitting (the situation where the algorithm fails to fit additional data or predict future observations reliably and its predictability is limited only on data similar to the original training set) even for relatively small sample.

Shin et al. (Shin, Taik, & Kim, 2005) and Min and Lee (Min & Lee, 2005) used SVM to predict bankruptcy for South Korean companies They came up with the conclusion that this method significantly outperformed discriminant analysis, Logistic regression and Neural Networks. Hui and Sun (Hui & Sun, 2006) adopted an SVM model to do empirical study on FDP for Chinese listed companies, and reached a similar conclusion. SVM approach has been introduced to several financial applications such as credit rating, time series prediction, and insurance claim fraud detection (Kewat, Sharma, Singh, & Itare, 2017), (Damrongsakmethee & Neagoe, 2017).

Tian and Deng (Tian & Deng, 2004) in their research, concluded that SVMs performed well in comparison with neural networks, genetic algorithms and classification trees in credit scoring. However, the learning algorithm may be inefficient and SVMs may be difficult to implement as a large number of parameters is required. In addition, according to Shawe and Cristianini (Shawe-Taylor & Cristianini, 2000), small training samples will result in overfitting, with poor generalization ability.

Chapter 7

Research objectives and research problem

7.1 Objectives

The aim of this thesis is to introduce an end to end credit scoring modelling process, starting from the initial steps of data selection, data preparation and data cleaning, to the final implementation of the model with real life data, in order to predict the probability that a borrower defaults on their obligations towards a financial institute. The process described in the next chapters follows a specific framework which is derived both from the bibliography and the theoretical part introduced in this Master's Thesis and also, from the experience of the Author as an analytics specialist during the last years. We have already explicitly introduced in the theoretical part of this thesis, the reasons why credit scoring is of great importance for a financial institute in order to be strong and as less vulnerable as possible towards risks. In simple words, if the decision makers of an organization had the ability to predict exactly the probability of default an applicant for a loan or funding may have, then the organization would have zero risk towards their investment as long as the defaulters would be recognized with perfect accuracy in their initial stage of application. But in real life, the exact calculation of the aforementioned probability is impossible, therefore an accurate predictive model would be a great asset for the financial institute and especially the risk management team, a "weapon" the organization would hold towards the *"battle of financial survival"* and risk minimization. A full detailed classification modelling process was considered very important to analyze from the Author of this Master thesis and this business need in addition to the research questions that follow generated the idea for this.

7.2 Research Questions

We try to give answers in three different research questions:

1. What is the process of constructing a classification model for credit risk scoring?
2. What incremental value does a classification model for predicting the probability of default add to a financial institution?
3. Which classification algorithm suits better for the purpose of credit scoring calculation and risk assessment?

Additionally, we will put under investigation and comparison four different classification models for the prediction of default and conclude in the model that best fits our objectives arguing towards the corresponding selection reasons.

7.3 Methodology

For the purpose of the analysis performed in this Master Thesis, an open source, free, large dataset consisting of more than 800 thousand different loan applications and 70 different quantitative and qualitative variables is selected from the archive of one of the biggest P2P and B2B lending institute located in US. The initial data are analyzed from the author of this thesis in order to conclude into a final dataset with less 300 thousand observations and less than 25 significant variables. Those metrics were then analyzed deeper through descriptive statistics and correlations. The final dataset is divided into training and test subset using k-fold cross validation method. in order to provide the models with the best possible datasets for training. Finally, 4 different classification models are implemented with the use of R statistical language and IBM SPSS Modeler and their results were tested on the aforementioned test sets. Then, findings and prediction results of all four models are put under comparison with the use of well-known accuracy measures provided by the bibliography. Lastly, we argue towards the incremental value of those models in a business perspective in order to support our argument that credit scoring is a great tool for minimizing credit risk.

In the following sections of this chapter, we will explicitly introduce an end to end, complete analysis of every single step taking place during the development of a credit risk model and generally, the process of creating and implementing a model, from data collection, data understanding, cleaning and preparation and major or minor issues may arise during the development of a predictive model to the implementation and evaluation of the model in terms of decision making. The steps that will be introduced should be considered as a general framework regarding a complete analytics process not only in credit risk modelling, but also in any aspect and sector of risk analysis and algorithmic implementation. The vast majority of every analytics project follows this guideline, with minor differences driven by the scope of every single analysis in terms of the results the analyst wishes to come up with. The author of this Master Thesis has used not only comprehensive literature to introduce this framework but also his experience from working as an analyst in a Customer Analytics department in one of the biggest enterprises in Greece.

7.3.1 Understanding the problem

The first step on every modelling process is to understand the problem in terms of business utility. Why an algorithm should be implemented and what incremental power this algorithm will add to the business in terms of profit maximization or risk minimization is the first question a risk manager should ask himself. In the broad world of banking and investment sector this question has a very simple answer; the need for better predicting a lender's future behavior towards his or her obligation to the funding institution is of vital importance in order the institution to avoid situations were mass defaults with catastrophic results "will hit the door". Though it is very important for every investing institution to better assess their risk exposure. And this is the part where Machine learning and statistical methods get into the game. The better the prediction model the institution implements when trying to predict the future behavior of the lender, the smallest the risk exposure towards a possible economical retention or maybe, bankruptcy.

7.3.2 Data Selection

Data selection is the next step in the long way until the extraction of a well fitted classification model. The types of data the analyst should collect in order to use on their models are highly related with the business scope of the analysis. In general, historical data held in organizations data warehouses, are provided by the Business Intelligence department to the analysts. In most cases those data come in raw or unstructured format and the analyst should go through a data preparation process which normally contains data cleaning, data manipulation and descriptive statistics in order to finally come up with a decision on which variables will hold his explanatory predictors' list.

In the next section, data preprocessing techniques and common challenges an analyst shall face during the modelling process are introduced.

7.3.3 Data Preprocessing and Data Cleaning

7.3.3.1 Missing Values

In real world data, there are instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. Handling the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. Let us look at different ways of imputing the missing values.

- ***Deleting rows with missing values***

This method is commonly used to handle the null values. It is broadly known as *Listwise Deletion* (Soley-Bori, 2013). Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output. The biggest advantage of deleting rows with missing values is that the analyst gains on robustness and higher accuracy. The major disadvantage is that we may loss significant observations, especially when taking about default prediction, where in most cases, we have very few default class observations and maybe a raw deletion will result in losing some of those few -but very valuable for the algorithm- observations.

- ***Deleting columns with many missing values***

Another option an analyst has to deal with missing values, is to drop all variables that have more than a certain percent of missing values. For example, he may choose to drop all columns with more than 50% of missing values. Those columns will not add value on the analysis, as the algorithm will "face significant problems" in evaluating those variables with such diverse number of available information compared to the other explanatory variables with no or minor number of nulls. Deleting columns is a decision the analyst should take also considering their unique business case and scope of analysis. In other words, if a variable has lots of missing observations, thus is significant in terms of business decision, then he shall may search for an imputation method (replacing missing values) rather than completely dropping a sense-making variable from the analysis.

- ***Imputing Missing values with Mean/Median/Mode value***

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values (Bennett, 2001). This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training (Bennett, 2001). Another way is to approximate it with the deviation of neighboring values. This works better if the data is linear. The limitation of these method comes in terms of leading to biased estimates of variance and covariance and thus, should be avoided (Soley-Bori, 2013).

- ***Conditional Mean Imputation – Regression of missing values***

This approach involves developing a regression equation based on the complete subject data for a given variable, treating it as an 'outcome' and using all other relevant variables as predictors (Bennett, 2001). For observations where the 'outcome' is missing, the predicted values from the regression equation are used as replacements. This method has similar problems to the mean imputation method but these problems can be overcome by adding uncertainty, usually by weighting, to the imputation of 'outcome' so that the mean value is not always imputed.

As we will see in the next chapters, our dataset has more that 20 values with over 80% of missing values. Those variables will be excluded from our dataset. Additionally, variables with less than 10% of nulls will be imputed using either mean or median value depending on the nature of the particular variables.

7.3.3.2 Descriptive statistics

Descriptive statistics are commonly used to understand and assess all available metrics in terms of their nature, type, values, extremes and every information they may provide to the analyst to help him better understand the dataset. Therefore, descriptive statistics stand as a very important step during data understanding process and data preparation. Initially, the analyst starts with simple metrics on every continuous and categorical variable. Metrics such as minimum and maximum value, mean, mode and median, standard deviation and range are initially calculated in order to provide the analyst with better insights on the data. Furthermore, descriptive statistics also provide valuable information for the final decision of which variables should be implemented into the algorithm and which ones should be rejected from the analytical process. For example, variables with only one value through all observations or on contrary, variables with a huge number of unique values are not good explanatory indicators for a predictive model.

Additionally, descriptive statistics help towards the extraction of possible relationships between explanatory variables itself or between explanatory variables and the target variable. Correlations and dependencies are also put under observation before diving into the final step of modelling process. In fact, descriptive statistics is a step which – in most cases – hold the most of the total time of an analytics project and should be regarded as of great importance towards the final result of the classification model.

7.3.4 Splitting into Training and Test sets

In order to be able to measure a classification model’s performance, we should initially provide the algorithms with the “opportunity to learn”. In other words, in order the model to be able to predict the final outcome of the target variable, it should be trained with a partition of all available data and after that, implement the implicit derived classification rules to the rest observations of provided data through which the efficiency and robustness of the model will be assessed.

Therefore, in every classification model, it is needed to divide the original data set into at least two parts; the training set (observations through which the model is trained) and test set (on which the classifiers can be scored). There are many different methods to split the original data set, and to create the testing data set. In this section, we will describe the three most common methods used in practice.

The first method is a simple, random separation into training and testing data sets. In the first step, the classifier is trained based on the training set. In the second step, a trained classifier is scored on the testing data set. The separation ratio of original data set is usually between 70% and 80% for training data and remaining 30% or 20% for testing data with variations according to the different business perspective of the classification process. The second method is slightly different and literately new compared to the original separation in training and test set. We can separate the original data set into three parts - training, validation and test sets. The most common split ratio is 65% for the training set, 25% for the test set and the rest 10% for the validation set.

According to the 3-sets variation split, after the classifier been trained, the validation data set is used for model fine-tuning. Salzberg (1997), argues towards the need of the validation data set as it helps to improve the classifiers out of training data set accuracy. Some classifiers might have near perfect accuracy based on the training data. However, these classifiers might perform then very poorly on the testing data (Salzberg, 1997). Therefore, the validation data set is used to fine-tune the classifier before being scored on the testing data set.

The last method used in practice and the one we have chosen to implement in this Master Thesis is called *k-fold cross-validation*. According to Salzberg (1997) and Huang et al. (2007), this approach provides valid and robust classification results compared to the other two methods. In the next paragraph, we will introduce in a detailed manner how k-fold validation method works.

In the k-fold cross-validation method the original data set is randomly divided into k subsets. Each of the k subsets is used as testing data set in one of the k iterations. The remaining k-1 subsets are used for model training and fine-tuning. This approach minimizes the impact of data dependency or in other words *"data overfitting"* is avoided. Thus, the risk that the performance of a classifier depends on the choice of testing set is minimized because the classifier is scored sequentially on the whole data set (Huang & Chen, 2007).

7.3.5 Handling imbalanced observations

A very often situation that an analyst has to deal with is the high degree of imbalanced observations through a dataset. Imbalanced datasets are those were the values of the target variable are highly skewed towards one category. The minority class instances often, as in our case, contains the information of interest and thus it is important to correctly classify them instead of predicting the majority class observations. In our case, that means, to predict the actual defaulters with the highest possible accuracy, rather than finding those applicants that will eventually fulfil their obligations. For a financial institute, identifying possible defaulters will save money and minimize risk on the long run. Therefore, it is much worst to mispredict a rare event than a more common one in the sense of consequences.

Misclassification of minority class instances are more likely when dealing with highly imbalanced datasets. Many classification algorithms use overall accuracy to optimize and this will make the prediction performance of the minority class worse, as stated in Ertekin et al (2007).

Among all possible solutions to the problems with imbalanced data, two commonly used methods are sampling based and cost function based. According to Sandberg (2017), sampling based methods can be divided into three approaches: oversampling, where more observations are added to the minority class, often by replicating existing observations; under sampling where observations from the majority class are removed; and also, a mix of the two have been considered. One example of a method which uses the latter approach is SMOTE (Synthetic Minority Over-Sampling Technique), but instead of replicating minority class instances a new minority class data observation are constructed using an algorithm, that borrows information from neighboring data points. Chawla (2002) argues, that SMOTE method overpasses the problem of minority class overfitting as may happen with replication of minority categories' observations.

7.3.5.1 SMOTE method

On their research on resampling methods for manipulating imbalanced datasets, Chawla et al. (2002) introduced SMOTE, a technique that generates synthetic instances examples in order to handle problems with overfitting and at the same time make the decision region bigger for the minority class. SMOTE method works in the following way; for each minority class observation, synthetic observations are created along the line segments by joining the observations nearest neighbors. In other words, for every two nearest minority class observations, SMOTE methods create one more observation with the same characteristics with the two neighbors to this synthetic point observations. In the algorithm, we specify n as the number of nearest neighbors and depending on the amount of extra observations required, some or all of these n neighbors are randomly chosen. For example, if we want to oversample 400%, two of the n-nearest neighbors are chosen and in each direction two samples of the minority class are generated (Chawla, 2002).

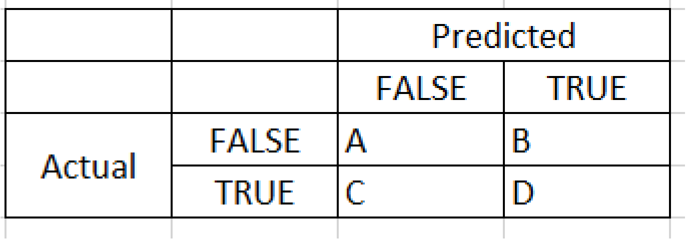
SMOTE technique is used in this Master Thesis for resampling the initial dataset minority class observations in order to increment the sensitivity of implemented classification algorithms.

7.3.6 Evaluation of the Model

7.3.6.1 Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. It is a technique for summarizing the performance of a classification algorithm (Kohavi & Provost, 1998). Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give the analyst a better idea of what a classification model is getting right and what types of errors it is making.

The main problem with classification accuracy is that it hides the detail needed to better understand the performance of the model. There are two examples where an analyst is more likely to encounter this problem; First, when the target variable has more than 2 classes, a classification model may give an accuracy of 80%, but this accuracy does not denote if all classes are being predicted equally well or whether one or two classes are being neglected by the model. Second, if the dataset does not have an even number of classes (the problem of imbalance), the model may achieve accuracy of 90% or more, but this is not a good score if 90 records for every 100 belong to one class and the model is achieving this score by always predicting the most common class value. The following table shows the confusion matrix for a two-category classifier:



		Predicted	
		FALSE	TRUE
Actual	FALSE	A	B
	TRUE	C	D

Letters A, B, C and D in the confusion matrix have the following meaning:

- A is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive (False positive),
- c is the number of incorrect of predictions that an instance negative (False Negative), and
- d is the number of correct predictions that an instance is positive.

The accuracy of the model (the percentage of correct predicted observations) is determined using the equation:

$$AC = \frac{A + D}{A + B + C + D}$$

The problem with model accuracy is that it cannot define in a proper way the true predictive ability of the model as it may depend solely on the prediction of the values of the largest class. Two other metrics derived from the confusion matrix are more appropriate and explicitly indicative for the evaluation of the model. Those metrics are True Positive (TP) and True Negative (TN) percentages (Kohavi & Provost, 1998). Recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the following equation:

$$TP = \frac{D}{C + D}$$

Recall shows the percentage of correct predictions that an instance is positive divided by the actual positive observations. on the other hand, TP is the proportion of negatives cases that were classified correctly. True Negative percentage is calculated using the following equation:

$$TN = \frac{A}{A+B}$$

Accuracy, recall and True Negative proportions are used to evaluate classification algorithm's performance. Additional to Confusion Matrix, Area Under Curve (AUC) and Receiver Operating Characteristics (ROC) are also two other evaluation metrics that have been used in this Master Thesis and introduced in the next section.

7.3.6.2 Area Under Curve and Receiver Operating Characteristics

A receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance. According to Swetz (2000), ROC graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers. Fowcett argues that ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems (Fawcett, 2005). ROC graphs are two-dimensional graphs in which True Positive rate is plotted on the Y axis and False Positive rate is plotted on the X axis. A ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives) (Fawcett, 2005). A demonstration of ROC curve is displayed in **(Figure 12).**

Several points in ROC space are important to note. The lower left point (0, 0) represents the strategy of never issuing a positive classification; such a classifier commits no false positive errors but also gains no true positives. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1, 1). The point (0, 1) represents perfect classification. One point in ROC space is better than another if it is left and down (True Positive rate is higher and False Positive rate is lower, or both) compared to the first one. Fawcett points that classifiers who appear on the left-hand side of a ROC graph, near the X axis, may be thought of as *''conservative''*, meaning that they make positive classifications only with strong evidence so they make few false positive errors, but they often have low true positive rates as well (Fawcett, 2005). On the other hand, classifiers on the upper right-hand side of a ROC graph may be thought of as *‘‘liberal’’*: they make positive classifications with weak evidence so they classify nearly all positives correctly, but they often have high false positive rates (Fawcett, 2005).

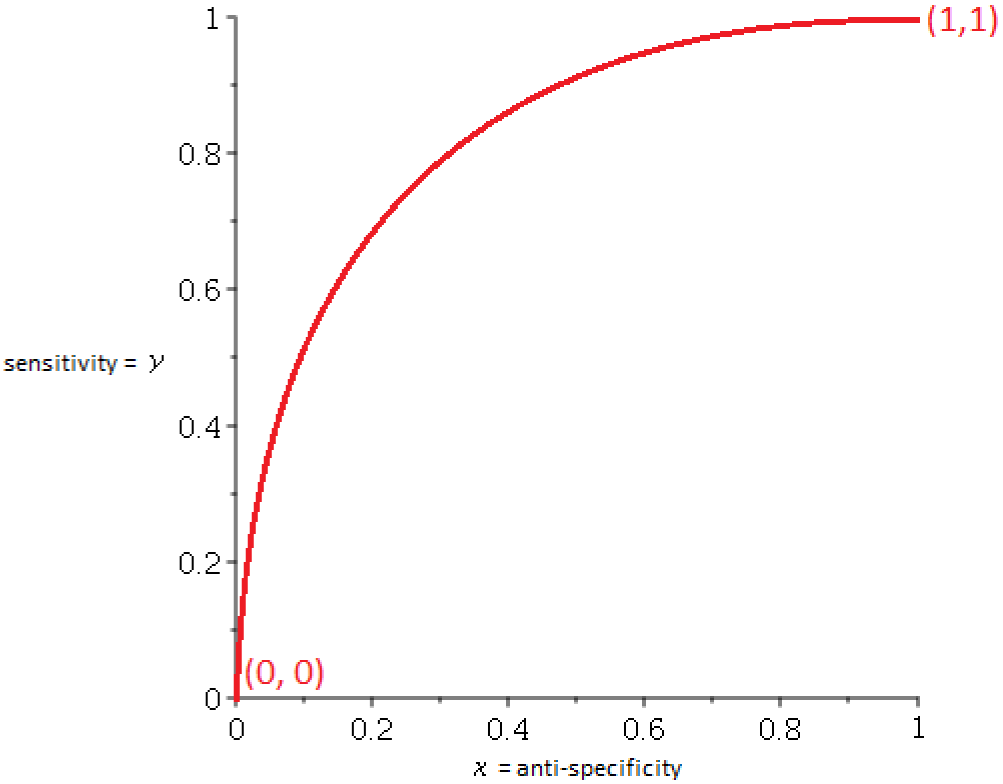


Figure 12: Receiver Operating Characteristics (ROC) Curve

Another metric commonly used to compare classifiers and also derived from the ROC curve is Area Under Curve (AUC). The AUC of a classifier equals the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Bradley, 1997). In other words, AUC displays the probability that a model will be able to distinguish between a positive class observation and a negative class observation. The higher the AUC, the better the performance of a classification model. A model with AUC equals 50% has no predictive power, as it distinguishes randomly between positive and negative classes of a binary outcome, while AUC = 1 corresponds to a “perfect” model which has the ability to undoubtedly recognize and classify an observation into the right class. AUC and ROC as well as Confusion Matrix are the three metrics which are introduced in this Master Thesis as model performance indicators.

Chapter 8
Modelling

8.1 Introduction of the dataset

The dataset used for the purpose of the analysis in this Master Thesis is provided by Lending Club Corporation and downloaded through the official webpage of the organization after an official registration. Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. According to Reuters (2014), Lending Club is the world's largest peer-to-peer lending platform. The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015. Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

The dataset includes loan information from an eight-year period, between 2007 and 2015. It consists of 74 different variables and over 850000 observations – loan applications. There are both continuous and categorical variables in the dataset. The target variable is *"loan status"* and this variable is modelled in this Master Thesis in terms of predicting the final loan status. The 73 explanatory variables are explicitly introduced and analyzed in the following chapters in order the author of this Master Thesis to access their predictability power and eventually choose which ones will be used as explanatory variables for the purpose of the analysis. A complete table of all variables and their meaning can be found in the Appendix of this Master Thesis.

8.2 Variable selection

In the previous chapter, we have introduced a variety of methods through bibliography for handling missing values. Cleaning our dataset in terms of missing values is the first step in our data preparation process in order to come up with the final dataset which will finally be used for the classification modelling process. For the purpose of the analysis performed in this Master Thesis, deletion of variables with many (more than 30%) of missing values is applied. Additionally, imputation techniques will also be deployed in case of variables with less than 1% of missing values.

8.2.1 Columns with null values

Since the number of variables is quite big (74 variables) we first start by calculating the percentage of null values in each column of the dataset. The reason we are doing this is that if a variable has many blank records (more than 10%), then this variable will be excluded from the dataset.

We start by checking the number of missing (null or blank) values in the data set. For a more comprehensive and reader-friendly approach, the percentage of null values for each column will be plotted in descending order starting from variables with higher percentage of nulls **(Figure 13).**

Looking at the plot and the relative variable table, we can easily figure out the "quality" of each variable in the data set:

1. 19 variables have more than 10% of missing values. 16 of those exceed 97% and 18 variables have more than 80% of missing data.
2. 3 variables have 7 - 8% of missing values.
3. 52 variables have less than 1% of missing values, 45 of which with zero percent.

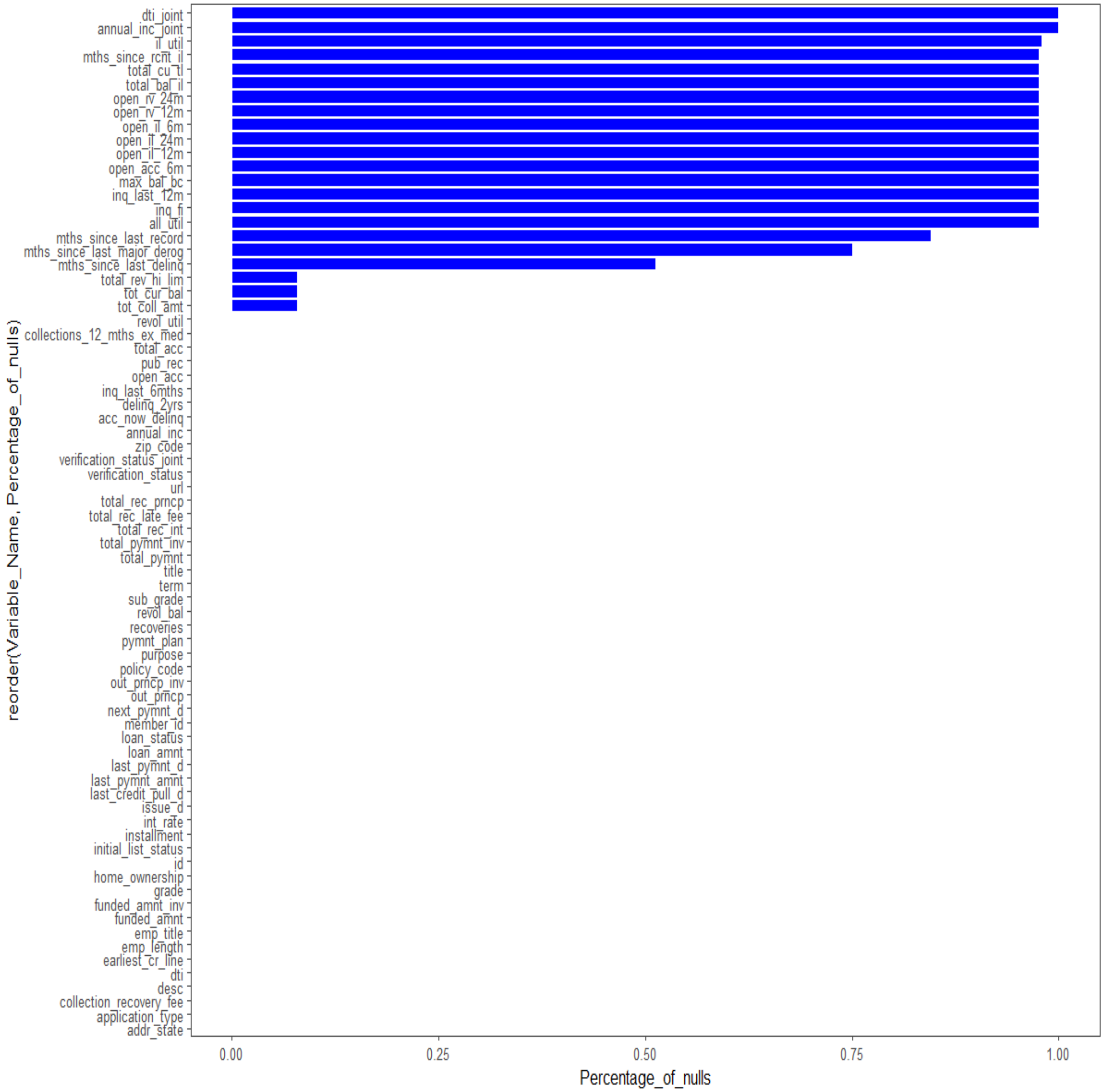
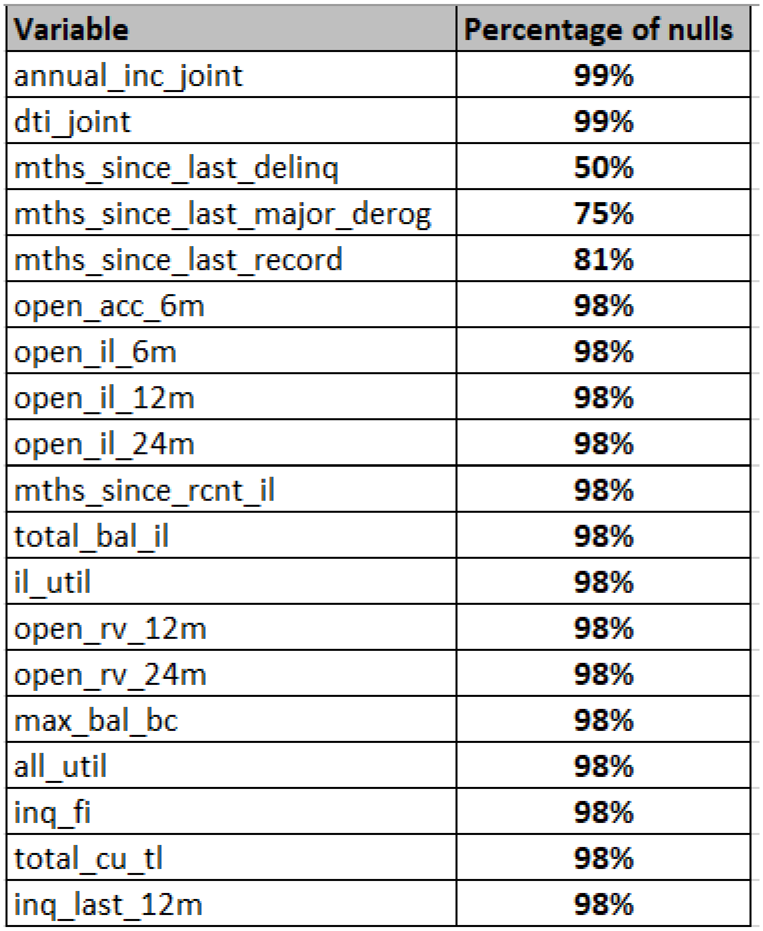


Figure 13: Percentage of null values through all available variables

In order to decide how to handle missing values, we should take a good look on those attributes not only in solid number but also in a business wise matter. Initially, we will exclude the nineteen variables with more than 10% of missing values from our dataset. Those variables (in descending order of the percentage of nulls) are shown in **(Table 1):**



Variable	Percentage of nulls
annual_inc_joint	**99%**
dti_joint	**99%**
mths_since_last_delinq	**50%**
mths_since_last_major_derog	**75%**
mths_since_last_record	**81%**
open_acc_6m	**98%**
open_il_6m	**98%**
open_il_12m	**98%**
open_il_24m	**98%**
mths_since_rcnt_il	**98%**
total_bal_il	**98%**
il_util	**98%**
open_rv_12m	**98%**
open_rv_24m	**98%**
max_bal_bc	**98%**
all_util	**98%**
inq_fi	**98%**
total_cu_tl	**98%**
inq_last_12m	**98%**

Table 1: Variables excluded due to high percentage of null values

Therefore, we now have a new dataset, with 55 variables that have less than 10% of missing values.

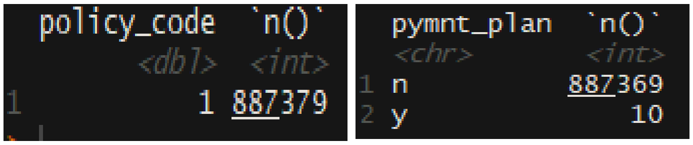
8.2.2 Removing redundant and future variables

The next step in our data preparation process is to remove variables which our redundant in terms of not providing useful information for the applicant and categorical variables with many different categories (for example *emp_title*) that algorithms are unable to handle as explanatory variables. Therefore:

- *id*, *member_id*, and *url* columns can be removed as they have unique values for the purpose of loan identification only.
- *zip code* column can be removed as it has only first three digits and that information can be obtained by the state column.
- We can also remove as we have already explained *emp_title* because it has more than 280000 unique emp_title and around 6% NA values and imputing them or deleting 50,000 rows from relevant data will lead to loss of data, hence, we will take out that column itself.
- *title column and verification_status_joint* can also be removed because they are redundant with purpose column which has fewer categories and *verification_status* respectively.

Additionally, we will remove variables that will not be present at the time of deciding whether to approve a loan or not and retain variables related to customer information and customer demographics. Therefore, variables such as *funded_amnt, funded_amnt_inv, issue_d, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_int, total_rec_late_fee, total_rec_prncp, recoveries, collection_recovery_fee, collection_12_mths_ex_med, acc_now_delinq, total_coll_amt, total_cur_bal, total_rev_hi_lim* are basically customer payment behavior parameters which will not be available during decision making and thus, the credit analyst will not be able to use them in order to calculate the credit score for a particular applicant.

Policy code and *pymnt_plan* should also be removed, as they have all values in one category and thus, cannot add incremental classification power to the algorithms (Figure 14):

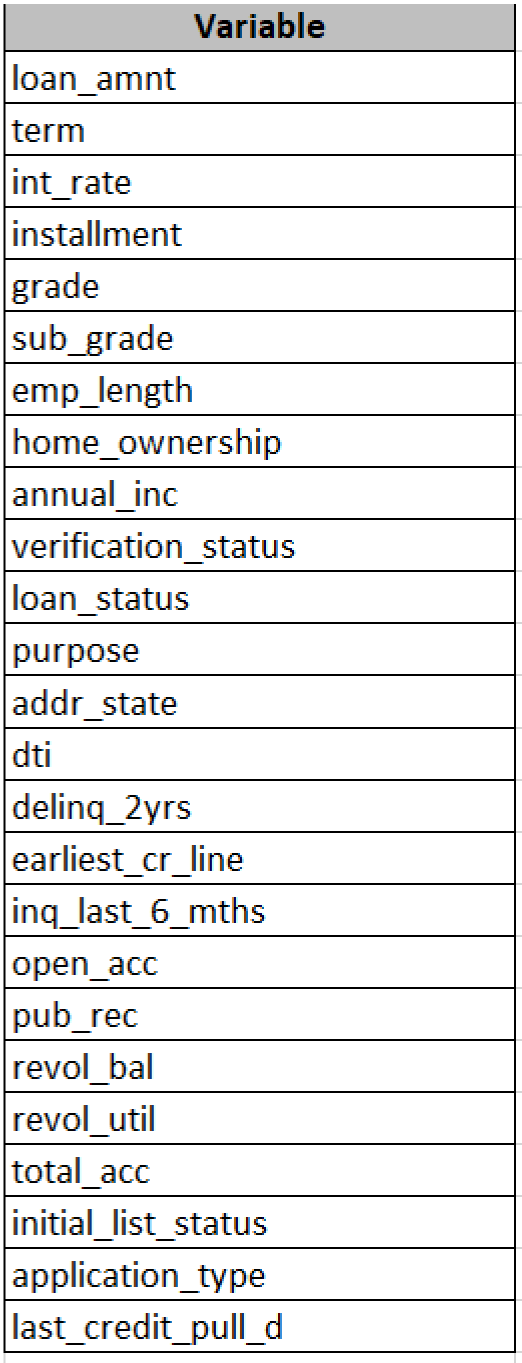


```
  policy_code  `n()`
        <dbl>  <int>
1           1 887379
```

```
  pymnt_plan   `n()`
  <chr>        <int>
1 n           887369
2 y               10
```

Figure 14: All values in one category

Hence, after those exclusions, we now have 25 remaining variables in our new dataset, 24 of which are explanatory variables and the last is our target variable *loan_status.* All remaining variables are shown in **(Table 2):**



Variable
loan_amnt
term
int_rate
installment
grade
sub_grade
emp_length
home_ownership
annual_inc
verification_status
loan_status
purpose
addr_state
dti
delinq_2yrs
earliest_cr_line
inq_last_6_mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
initial_list_status
application_type
last_credit_pull_d

Table 2: Variables included in the final dataset

8.3 Deriving new KPIs

For the purpose of our analysis, we will calculate some new key performance indicators, new variables from our initial data. The reason why we shall proceed in this step is that in many cases the need of grouping some of the continuous variables of a dataset is highly recommended in order to provide a classification algorithm with derived categorical variables from continuous variable with fewer values than the initial explanatory indicator may have.

The first metric will be derived by applicants' annual income. We will calculate a new variable called *income_category*, with 6 groups, from "Very Low" to "Very High" annual income. Additionally, and for the same reason, interest rate will derive a new metric called *"rate_category"*. This categorical variable will have 3 categories; "Low Rate, Medium Rate and High Rate. Finally, the last KPI which will be derived is *"Open_to_closed_accounts_ratio"*.

All those three new metrics are explicitly described in the following chapters where data exploration through descriptive statistics is performed.

8.4 Descriptive Statistics

In this section, we will investigate the characteristics of the majority of variables provided in the dataset – especially the most significant ones in terms of analytical interest – through descriptive statistics and diagrams in order to better assess and understand their nature. At first, descriptive statistics for the most significant explanatory variables will be introduced and after, we will go through a detailed analysis of the target variable (loan status) -the characteristics and attributes of this variable- in order to understand its nature for the purposes of our analysis.

Finally, we will search for correlations between the target variable and the most significant explanatory indicators in order to find possible relationships between them.

8.4.1 Descriptive Statistics on the explanatory variables

To begin with, we will take a look on the distribution of loan amounts provided by the Lending Club to the borrowers in order to get insights about the average loan amount that is approved. As we can see in **(Figure 15),** loan amounts' distribution is heavily divergent, with the mode value of loan amount to be 10.000 dollars – 80000 applications have been approved at that amount-.

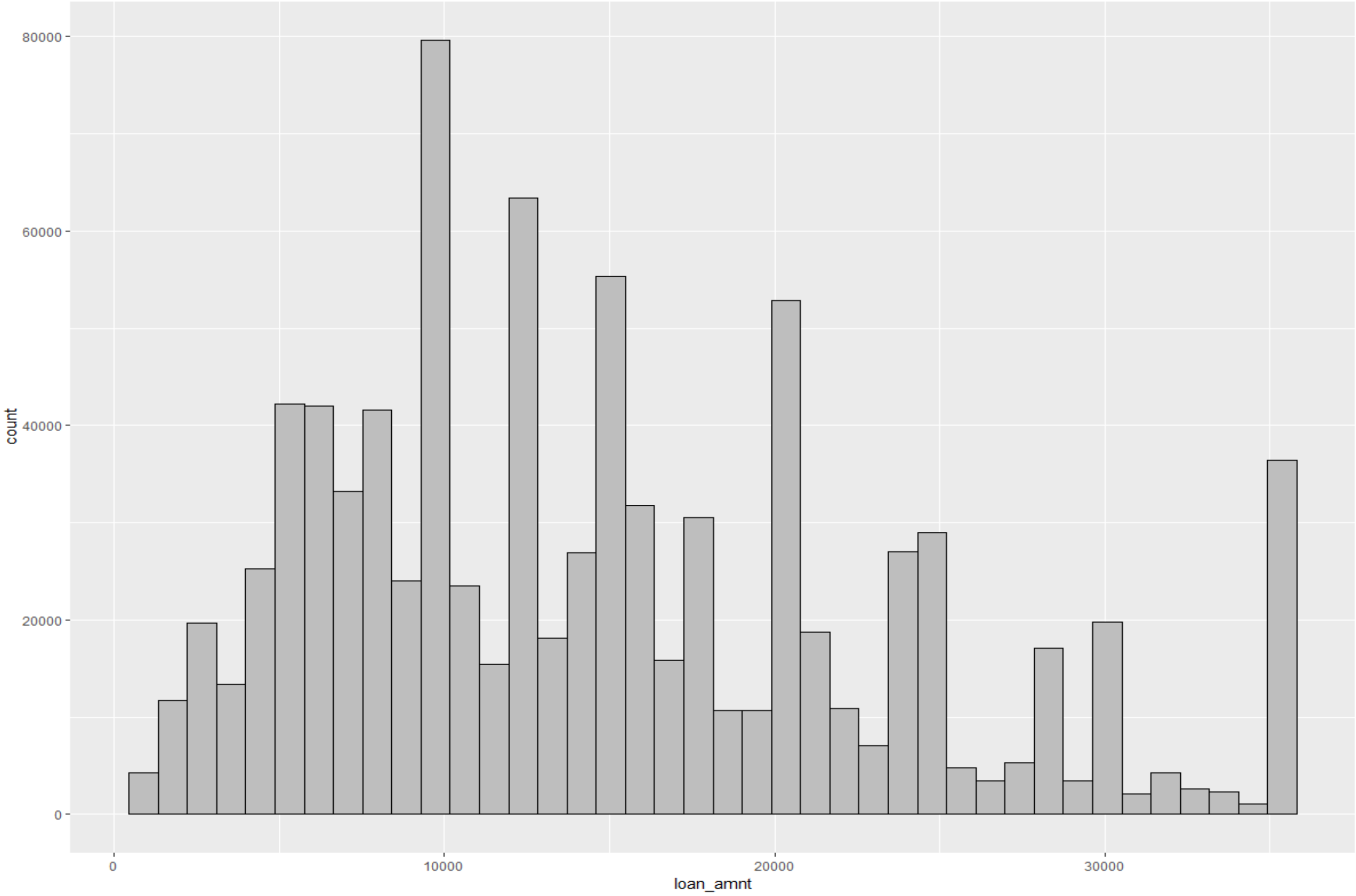
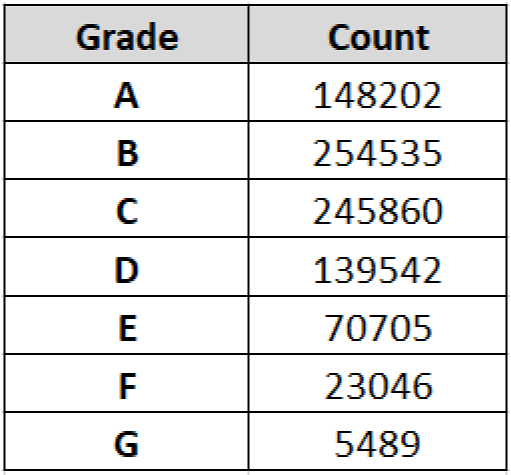


Figure 15: Distribution of Loan amount

There are also 60000 loans of 12000 dollars, 55000 loans of 20000 dollars. Furthermore, there is a significant total of 40000 loans with more than 35000 dollars approved to an applicant. The distribution of loan amounts is skewed towards smaller amounts – 447000 loans that have been approved are less than 13000 dollars-.

Next, we will look into the distribution of loans in term of their grade and subgrade. There are 7 different loan categories in terms of their grade in our dataset. The frequency table of loans, grouped together by their grade is represented in **(Table 3)** below:



Grade	Count
A	148202
B	254535
C	245860
D	139542
E	70705
F	23046
G	5489

Table 3:Frequency table of loans according to their grade

Loans of grade B and C are the most frequent, followed by loans with grade A and D. As a reminder, grade is assigned by Lending Club to every approved loan at the time of the approval. Loans are characterized as of grade A to grade G. The way the characterization is generated is not clear to the author of this Master Thesis and in order not to early drop this variable from the analysis, we assume that the grade of a loan is not arbitrary defined by LC experts. In addition, and under the same criteria, this assumption has been made also for the variable “subgrade”.

Let’s return on grade’s distribution. As we have already seen in **(Table 1),** the vast majority of loans are of grade C or better. In more detail, 648597 loans – 73% of the total 887379 - have grade A to C, 210247 loans (24%) are characterized with D and E and only 28535 (3%) are loans with very bad grades (F and G). The aforementioned details are summarized as a bar chart in **(Figure 16):**

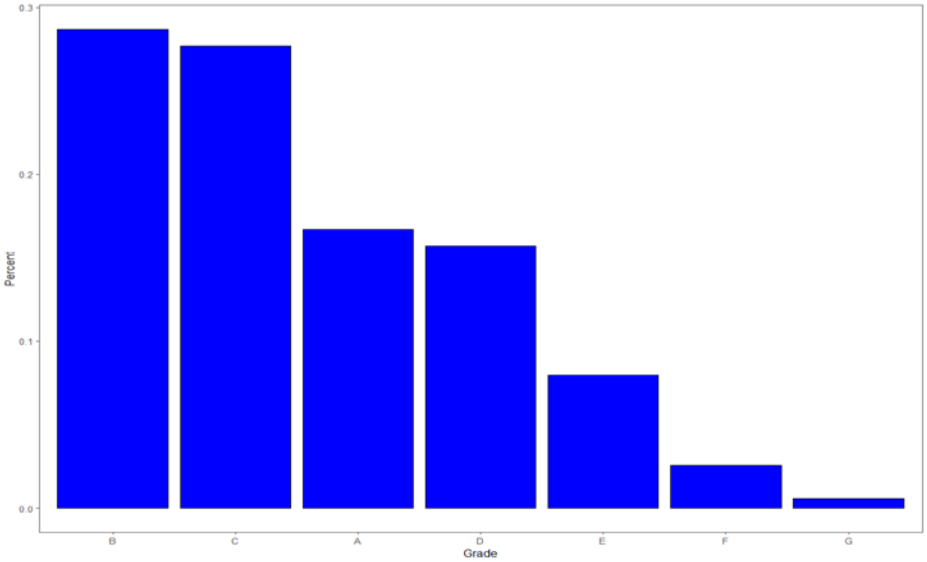


Figure 16: Bar Chart of the distribution of loans in terms of their grade

Exactly the same descriptive metrics are calculated for subgrade variable. Subgrade has 35 different categories starting from A1 and going down to G5 (A1: Best subgrade between all grades and G5: worst subgrade). The frequency table and the distribution graph of loans in terms of their subgrade is represented below **(Table 4), (Figure 17):**

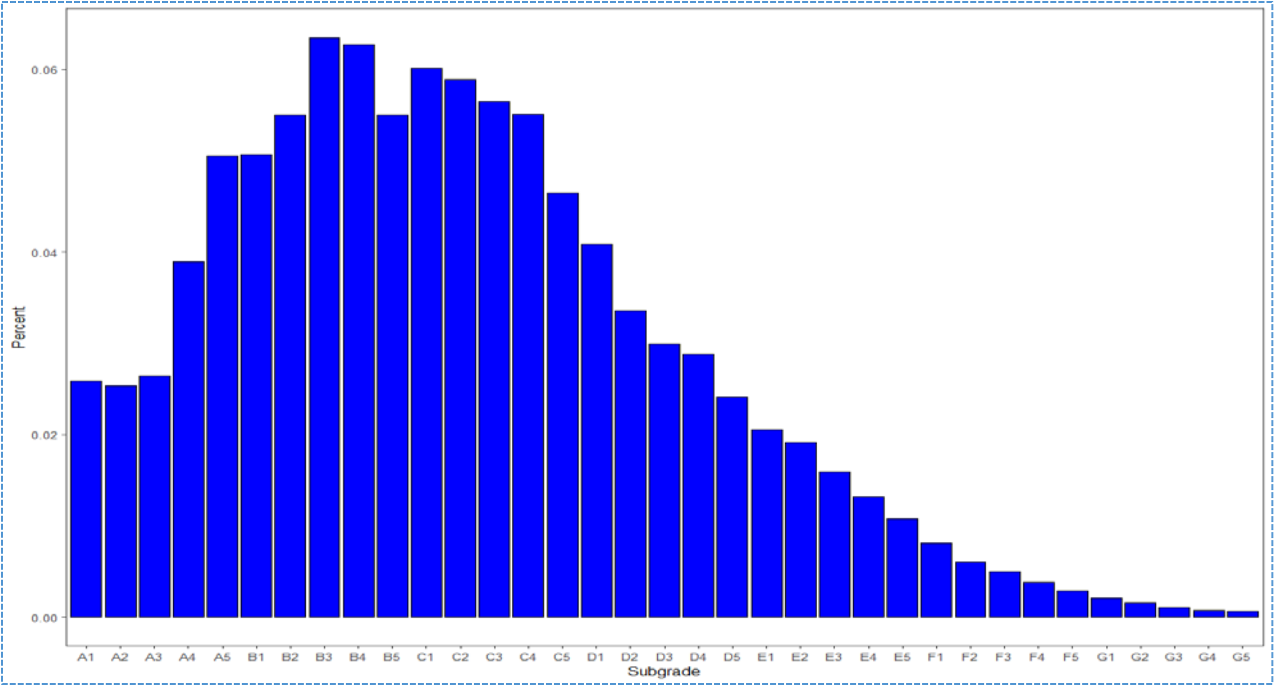
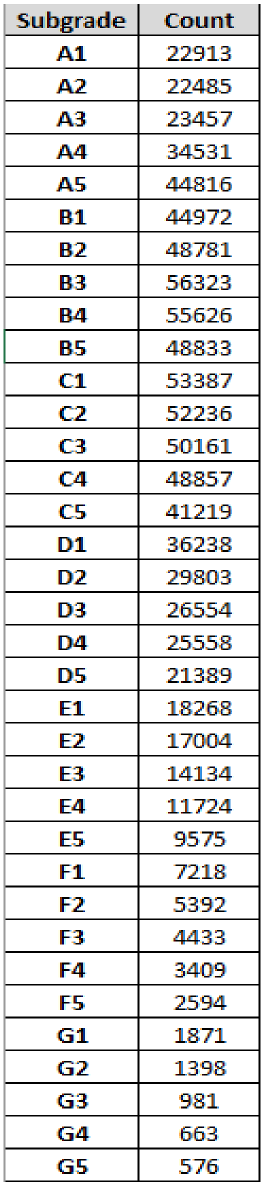


Figure 17: Distribution of loans in terms of their subgrade



Subgrade	Count
A1	22913
A2	22485
A3	23457
A4	34531
A5	44816
B1	44972
B2	48781
B3	56323
B4	55626
B5	48833
C1	53387
C2	52236
C3	50161
C4	48857
C5	41219
D1	36238
D2	29803
D3	26554
D4	25558
D5	21389
E1	18268
E2	17004
E3	14134
E4	11724
E5	9575
F1	7218
F2	5392
F3	4433
F4	3409
F5	2594
G1	1871
G2	1398
G3	981
G4	663
G5	576

Table 4: Frequency of loans in terms of subgrade

As it was expected, due to the skewness of grades, subgrades are also skewed to the left, towards better subgrades. The mode of subgrades is category "B3" with 56323 loans characterized as of that subgroup. Regarding the two “extremes”, there are 22913 “excellent applicants” (A1) and 576 “almost certain to default” loans (G5).

We will now investigate the distribution of loan amounts in terms of their grade. We will try to figure out if there is a significant differentiation between loan amounts and grades. In other words, if there is a specific grade that sums up the majority of loans into it. Therefore, we will construct a bar chart plotting total loans amount ever given in each grade category **(Figure 18):**

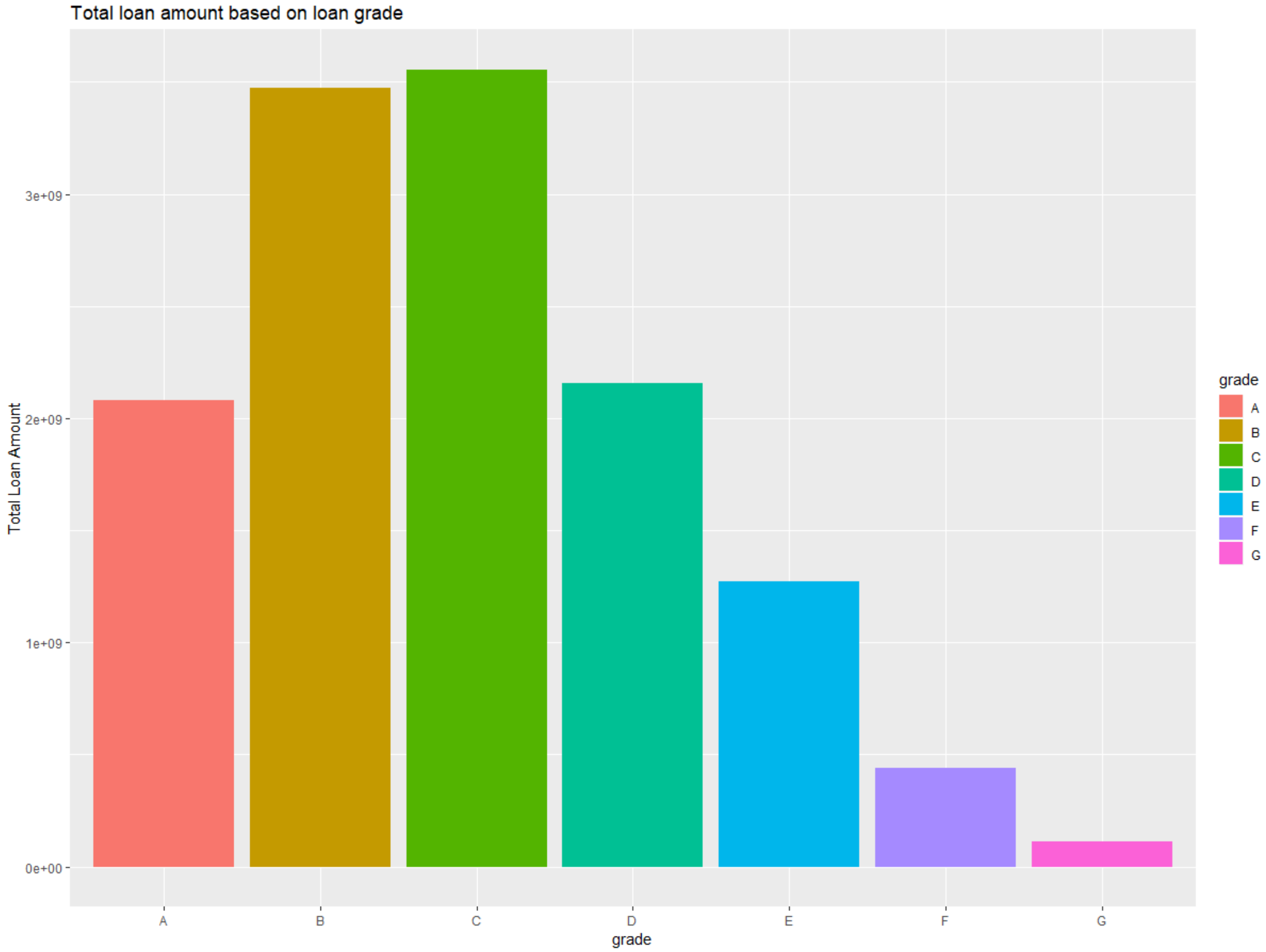


Figure 18:Total loans amount per grade category

As we can see, the greatest aggregated amount of money provided to borrowers by Lending Club, depicts on grades B and C. This could may be an indicator that grades B and C are loans that may provide a greater profit opportunity both for investors and borrowers.

Figure 19, shows the number of loans issued to borrowers per year. Loans were issued in an incremental manner, following a continuously increasing trend. Specifically, during the eight-year period between 2007 and 2015, the number of loans issued have almost been doubled (from 8 to 16 thousand loans) indicating a possible recovery to the U.S economy.

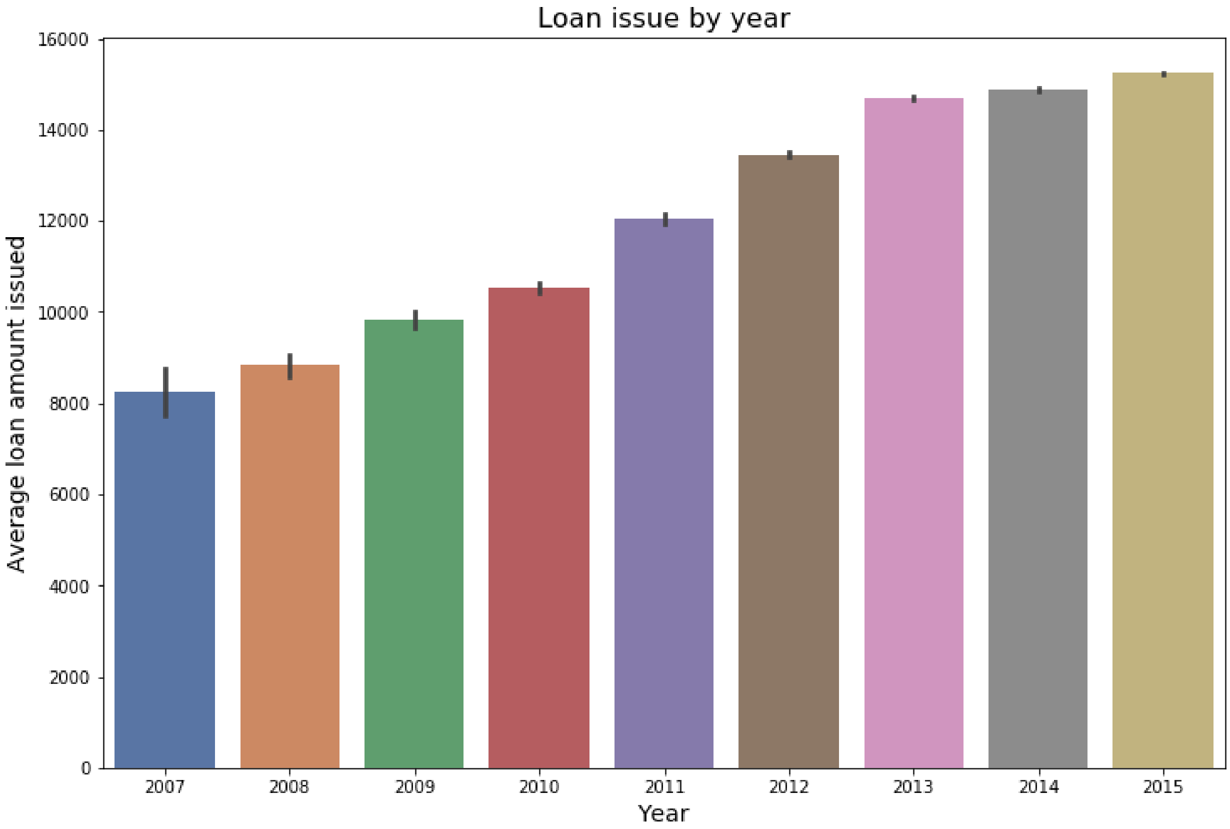


Figure 19: Number of loans issued per year

In **(Figure 20),** the distribution of loans in terms of their duration is represented. As we can see, there are two possible options for the borrowers when applying for a loan; choosing between 36 or 60 payments (installments). The majority of loans are given on 36-month term (almost 80%); this may indicate an initial intention from the borrowers' side, to complete their obligations the soonest possible.

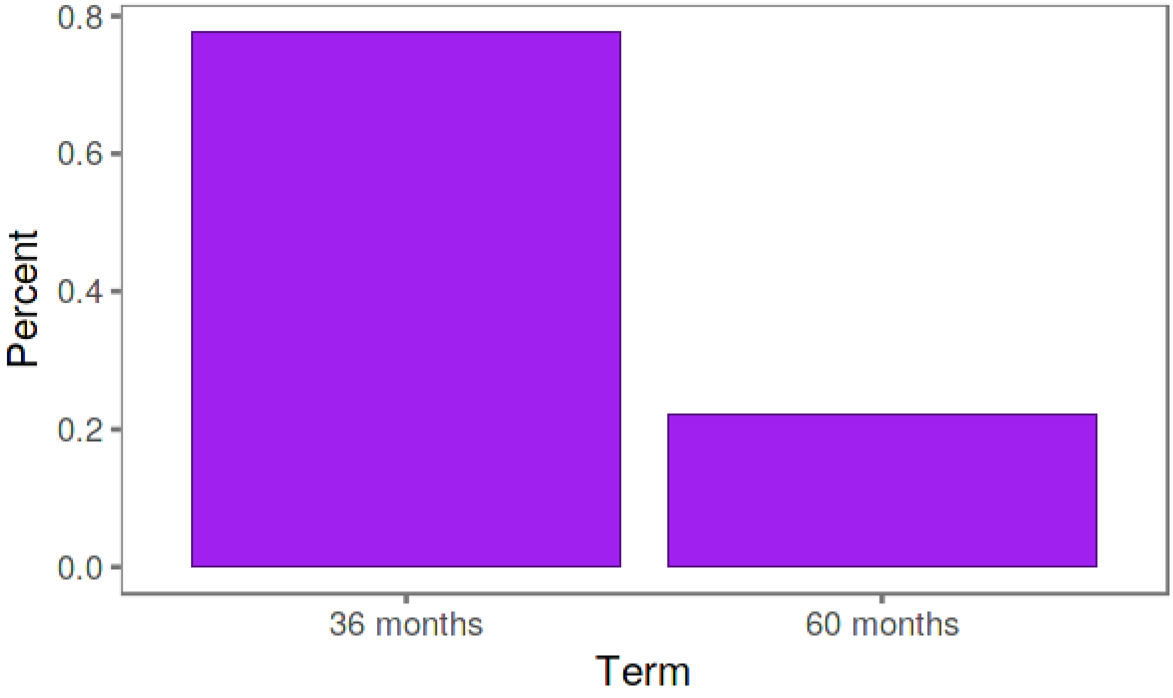


Figure 20:Payment Duration distribution

The employment length is another information provided by the borrowers when applying for a loan. Investors may consider this information as an important indicator of a borrower's ability two fulfill their obligations until the end and therefore, they may prefer borrowers with as much years of occupation as possible. Those thoughts seem to be confirmed if we take a look on the distribution of employment length **(Figure 21).** As we can see, the majority of borrowers have more than 10 years of employment length (almost 30%). Very interesting is the fact that the next most preferable occupational length is 2 years, followed by a significant 8% of applicants with less than 1 year of experience. The author of this Master Thesis considers employment length as a very interesting variable and in what comes next, he will try to investigate further relationships between this variable and other explanatory indicators.

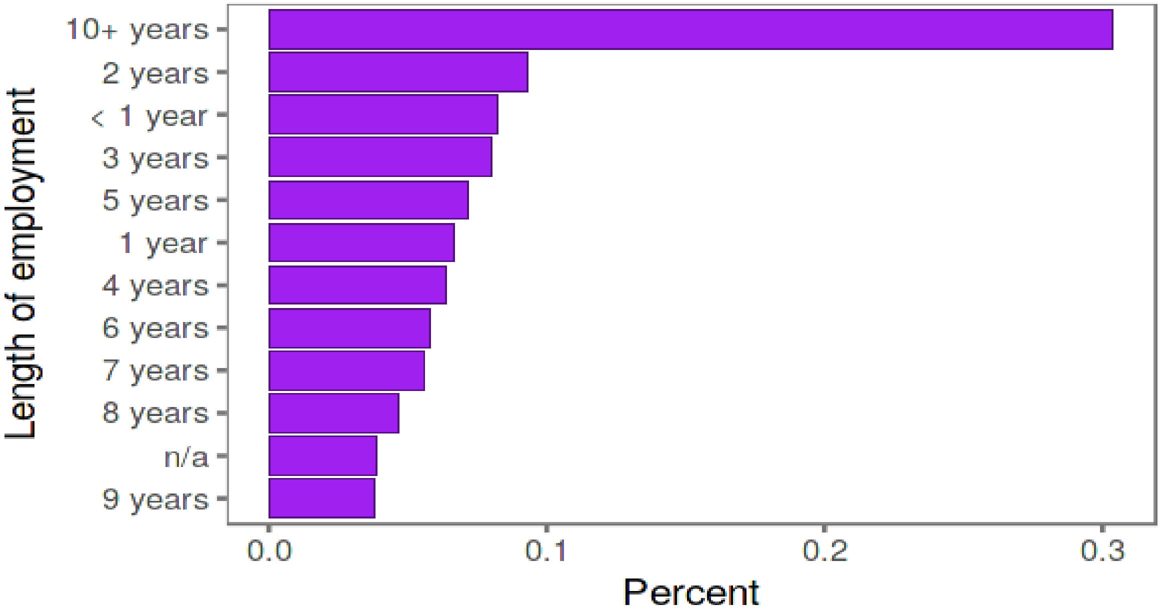


Figure 21: Distribution of Employment Length

(Figure 22), represents the home ownership status distribution of the 887000 applicants. Almost 90% of all applicants have either a mortgage or living in a house in rent. There is also a 10% of applicants who own a house without any mortgage. The Home Ownership variable will be further investigated in terms of correlation with other explanatory variables, but also in relation with the loan status (target variable), as the author believes that there may be a relationship between home ownership and default rate.

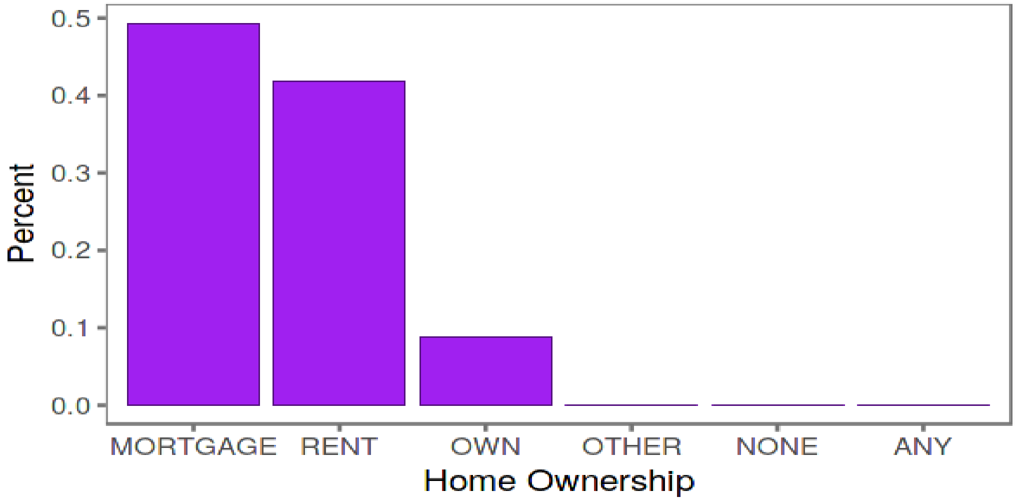


Figure 22: Home Ownership distribution of applicants

In terms of the purpose the applicant asked for a loan, the vast majority of borrowers asked for funding in order to consolidate an existing debt (60%). 20% of applicants asked for a loan in order to pay their obligations in credit cards, while another 10% is for improvements in their home, like renovation **(Figure 23):**

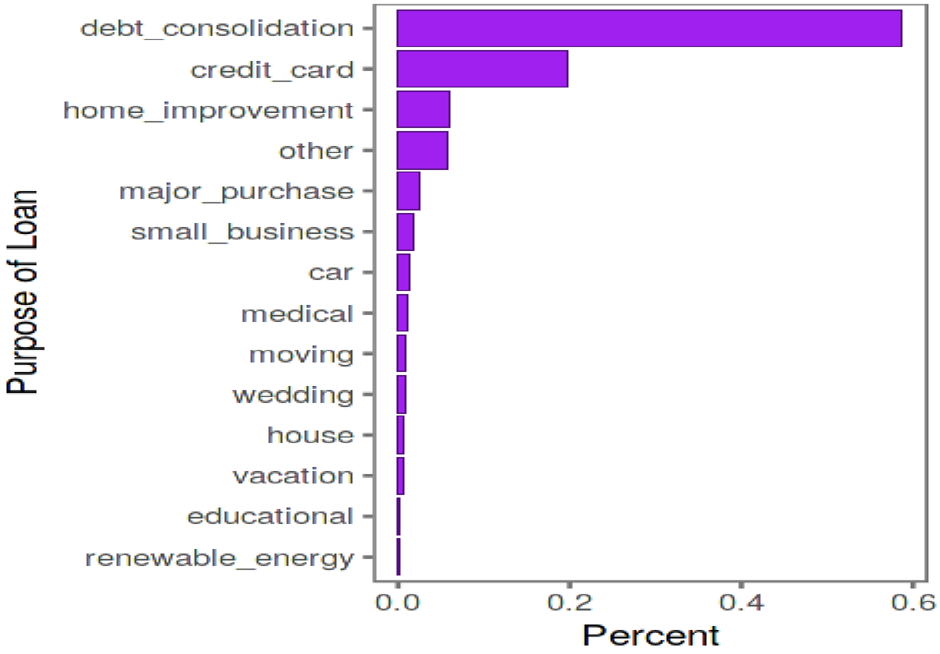
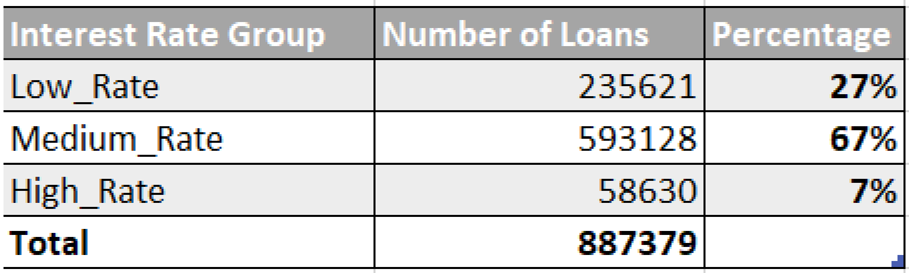


Figure 23:Purpose of loan

Another variable which may extract valuable information about the nature of loans and especially, the way investors decide on borrowing their money on a specific applicant is interest rate *(int_rate)*. The minimum interest rate is 5.32 percentage while the maximum is 28.99. The range is 24.67 indicating a very diverged dataset on how loans are evaluated in terms of risk exposure. In order to better assess loan applications in terms of their predefined interest rate, we will group this variable in 3 categories and then, we will calculate and demonstrate different comparative metrics between interest rate and other variables.

Though, from now on, loans with interest rate less than 10% will be characterized as “Low_Rate”, when loans with interest rate between 10% and 20% would be “Medium_Rate” loans. Finally, loans with more than 20% are characterized as “High_Rate” for the purposes of our analysis. The distribution of the new grouped loans is represented in **(Table 3):**



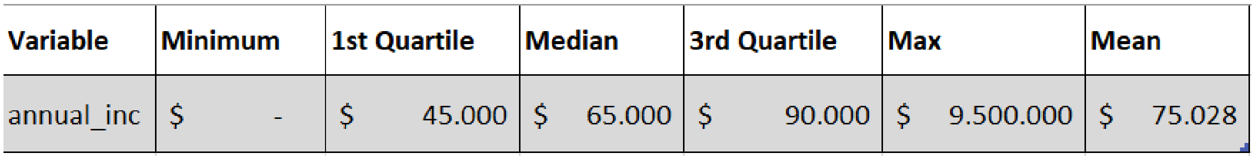
Interest Rate Group	Number of Loans	Percentage
Low_Rate	235621	**27%**
Medium_Rate	593128	**67%**
High_Rate	58630	**7%**
Total	**887379**	

Table 5:Distribution of loan applications per interest rate group

The vast majority of loans are of medium interest rate (67%), followed by loans with low interest rate (27%). The fact that the majority of approved applications are holding a return on capital between 10 and 20 percentage points reveals the risk appetite of investors; the higher the interest rate, the bigger the risk in terms of default but also, the greater the opportunity to gain more money from their investment.

Another variable that may be interesting to analyze in order to understand how it influences an investor's decision on funding a loan application, is the annual income of the borrower.

One may argue that the higher the annual income of an applicant, the more desirable this applicant may be for a potential investor to fund their loan. Annual income may be an indicator of the credit health of an applicant and a very useful metric for the calculation of the credit score of an individual. In **Table 4**, we can find detailed descriptive statistics of annual income *(annual_inc):*



Variable	Minimum	1st Quartile	Median	3rd Quartile	Max	Mean
annual_inc	$ -	$ 45.000	$ 65.000	$ 90.000	$ 9.500.000	$ 75.028

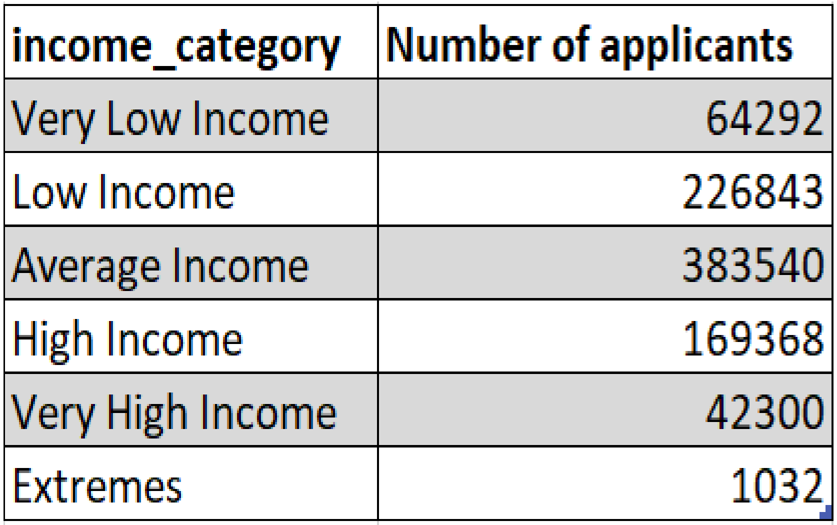
Table 6:Annual Income descriptive statistics

The minimum value of the variable is 0 dollars, when the maximum is an extreme of 9.5 million dollars. The average annual income of the applicants is 75 thousand dollars and the median is 65 thousand dollars. The 1st quartile is 45 thousand dollars and if consider the smaller median compared to the mean value, the distribution of the values is slightly skewed to the left. Interesting may be to dive deeply into the extreme values of this variable.

The fact that the maximum value is 9.5 million may be a problem for our analysis given that this outlier is extremely diverged from the rest dataset values. Therefore, we will take a closer look to the distribution of loans in terms of applicants' annual income.

We will group income values in 6 categories; "Very Low Income" for annual income less than 30000 $, "Low Income" for values between 30000 $ and 50000 $, "Average Income" for applicants with an annual income which is more than 50000$ but less than 90000$, "High Income" for values between 90000$ and 150000$, "Very High Income" if the observation is between 150000$ and 500000$ and "Extremes" for an annual income which is more than half million dollars. The new grouped values are gathered in a new variable called "*income_category*".

The aggregated results are demonstrated in **(Table 7):**



income_category	Number of applicants
Very Low Income	64292
Low Income	226843
Average Income	383540
High Income	169368
Very High Income	42300
Extremes	1032

Table 7: Applicants per annual income category

There were also 4 observations with null values. Those observations will later be excluded from the final dataset. The extremes are only 1 thousand applicants, demonstrating less than 0.2% of the total applications. Those extremes are also “candidates” for exclusion from our analysis in order to discard values that may affect our results and consequently are very few to be considered as significant values in our dataset.

The majority of applicants are of “Average Income” followed by “Low” and “High Income” borrowers. There are also less than 8% of applicants with “Very Low” or “Very High” annual income. What is revealed from the results of Table 5 is that investors prefer to borrow money on average income applicants maybe with a higher interest rate, a thought of the author of this Master Thesis the truth of which will be analyzed in the next chapters.

Now that we have analyzed and represented in detail the most significant explanatory variables of Lending Club dataset, in the next chapter, we will go through a detailed analysis of the target variable, loan status.

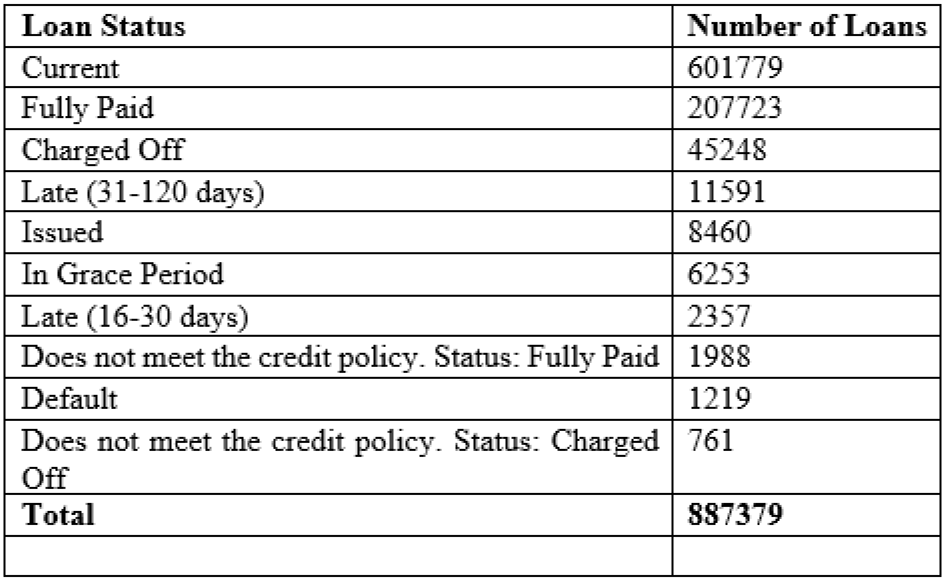
8.4.2 Descriptive statistics of the target variable

In this chapter, we will take a closer look on the different loan statuses in order to focus deeply on and further understand the target value. We will also argue towards the loan statuses that will be used during the modelling process. Not all categories are proper for use by the algorithms. The reason is that a credit analyst should take into consideration only loans with a final, not arbitrary loan status, in order to “help” the algorithm derive the best possible predictive results. In other words, loans with status “Current” for example are not suitable for use as the analyst has no knowledge on the final status of the corresponding loan; he cannot predict the future. Therefore, he should only take into consideration historical data with a well-known, final and not arbitrary loan status. More on the final selection of the target variable values will be discussed in this chapter.

Now, let’s take a deeper look into our target variable. As we have already described, the target variable is *“loan status”*. It describes the status of each specific loan in terms of whether the borrower fulfils their obligations towards the loan or not.

(Table 8) and **(Figure 24),** demonstrate the distribution of all loans issued between 2007 and 2015 per loan status. As we can see, the majority of loans are in status “Current”. This status indicates that a loan is currently paid in time by the borrower with no delays in payments. Additionally, there are 207723 loans that have already been fully paid by the borrowers and 8460 loans that have been issued during the time the data were provided. Issued loans will be excluded from our analysis. This decision has been made by the author of this Master Thesis, because Issued loans’ future status is not available by the time the data are obtained and thus, these loans cannot be categorized correctly.

Of course, we may assume that the amount of Bad Loans will increment during the maturity period of loans in status “Current”. But by the time the data are obtained, those loans are paid in time, thus, should be regarded as “Good”.



Loan Status	**Number of Loans**
Current	601779
Fully Paid	207723
Charged Off	45248
Late (31-120 days)	11591
Issued	8460
In Grace Period	6253
Late (16-30 days)	2357
Does not meet the credit policy. Status: Fully Paid	1988
Default	1219
Does not meet the credit policy. Status: Charged Off	761
Total	**887379**

Table 8:Distribution of loan status

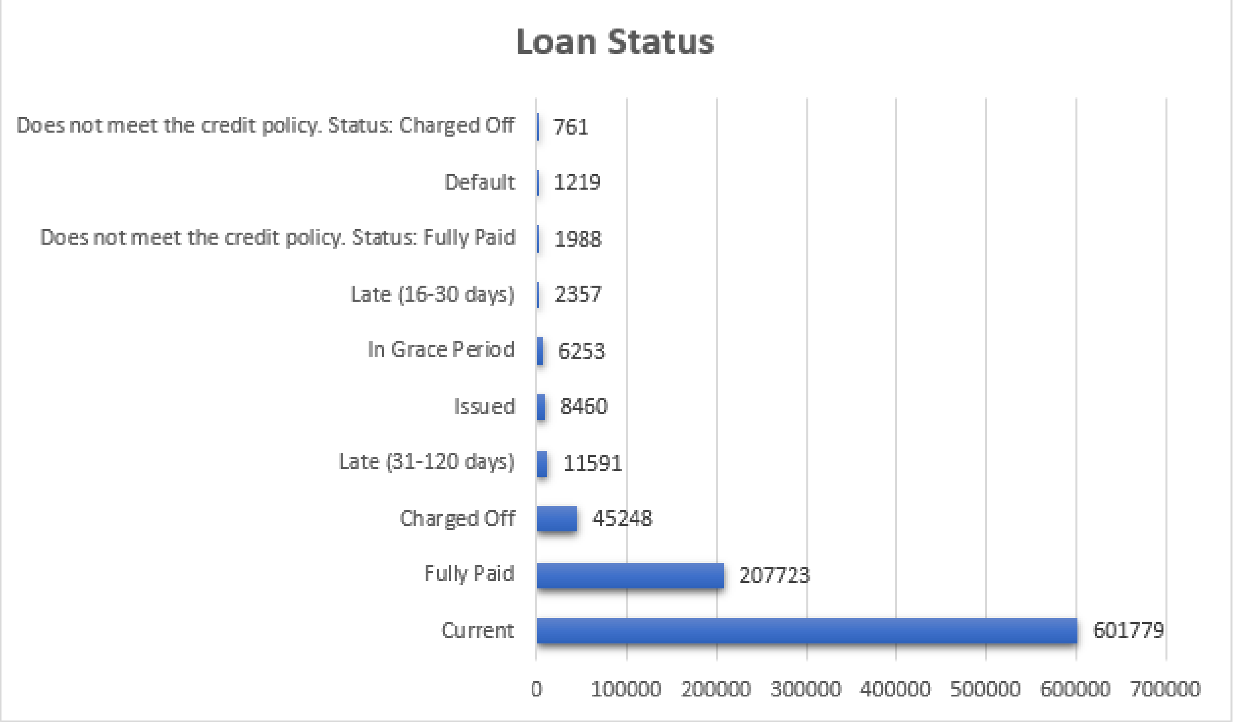


Figure 24: Bar chart of the distribution of loan status

The other five categories are characterized as "Bad Loans". Let's explain in detail each one of those categories:

- Status "Default" refers to loans that the borrower's defaults on their obligations, ie. borrower has failed to repay his loan. There are only 1219 loans in status default - less than 1% of total applications – indicating the extreme imbalance nature of the dataset.
- Status "Charged Off" refers to a debt that is deemed unlikely to be collected by the creditor because the borrower has become substantially delinquent after a period of time. A charge-off usually occurs when the creditor has deemed an outstanding debt is uncollectible; this typically follows 180 days or six months of non-payment. The creditor crosses off the consumer's debt as uncollectible and marks it on the consumer's credit report as a charge-off. There are 45248 loans – 5% of total observations that have been characterized as "Charged off". Those loans shall be considered as "Bad" for the purpose of our analysis.

- There also 6253 loans “In grace period”. A grace period is a period immediately after the deadline for an obligation during which a late fee, or other action that would have been taken as a result of failing to meet the deadline, is waived provided that the obligation is satisfied during the grace period. Grace periods can range from a number of minutes to a number of days or longer, and can apply in situations including arrival at a job, paying a bill, or meeting a government or legal requirement. In any case, when a loan ender a grace period, it is more likely to become “charged off” than current.
- Finally, there are also 13948 loans in delay of payment up to 4 months. These loans will also be regarded as bad ones.

Therefore, if we sum up the observations of the different loan statuses in each one of the two categories – Good and Bad loans, we come up with 66668 loans characterized as “Bad” and 876170 loans characterized as “Good”. The percentage of Bad loans over Good is 7.6%, which indicates a highly imbalanced dataset as we have already explained.

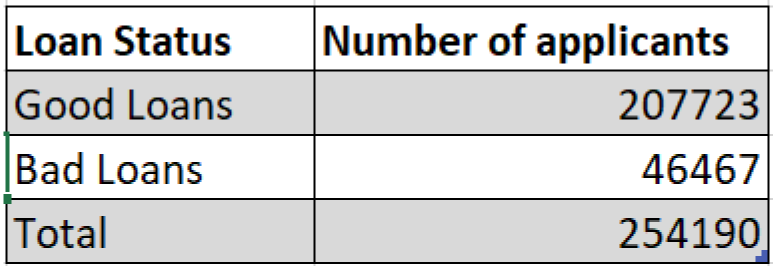
The proper definition and selection of the values of the target variable is of great importance as we have already described in the beginning of this section. When an applicant asks for a loan, the analyst takes into consideration the different attributes of the applicant (explanatory variables) in order to assign this applicant a credit score. Next, this credit score will be used as a credit worthiness indicator – for a given threshold decided by the institute that provides the loan – in order to decide if this loan should be granted or not.

The analyst uses historical data from past applicants with similar attributes. The applicant’s characteristics will be provided in a classification model which has already been trained with the available historical data. In order for the algorithm to give the better possible predictive accuracy, historical data should not be arbitrary concerning the target variable. In other words, the analyst should train the algorithm with data from loans with the greatest possible maturity; loans that have already been either fully paid or defaulted.

Though, for the purpose of our analysis, we will only keep observations either with status “Fully Paid” or with statuses “Charged Off” and “Defaulted”. “Fully Paid” loans will be considered as “Good Loans”, attached in applicants who have completed their obligations in time, whereas “Charged Off” and “Defaulted” should be considered as “Bad Loans”.

The reason why we categorize “Charged Off” loans as Bad ones, is that those loans are in fact ready for default and the creditor treats those as uncollectible as we have explained before.

Therefore, all other categories are excluded from our analysis. The new dataset consists of 254190 observations, the distribution of which in terms of their status is demonstrated in **(Figure 25):**



Loan Status	Number of applicants
Good Loans	207723
Bad Loans	46467
Total	254190

Figure 25:Final loan status distribution

Thus, we come up with a “new”, binary target variable, with values 0 and 1; zero for “Bad Loan” (Default or Charged Off) and 1 for “Good Loan” (Fully Paid). The new binary target variable is not highly imbalanced compared to the initial one, a fact that provides the algorithms with more incremental predictive power. Good Loans are 80% of total observations while Bad Loans are the rest 20%.

In what comes next, we will investigate our new target variable in terms of relationships with some of the already introduced explanatory variables. We will plot different explanatory variables in comparison with the target variable searching whether certain predictors affect loan status and additionally, we will search for correlations in order to prepare the ground for the implementation of our classification models.

8.4.3 Comparative metrics and correlations

At first, we will plot *"loan status"* towards the explanatory variable *"term"*, i.e. the agreed period for repaying the loan. As we have already described, an applicant has two options for repaying their loan; choosing between a 36- and 60-month period. A stack bar of loan status versus term is presented in **(Figure 26):**

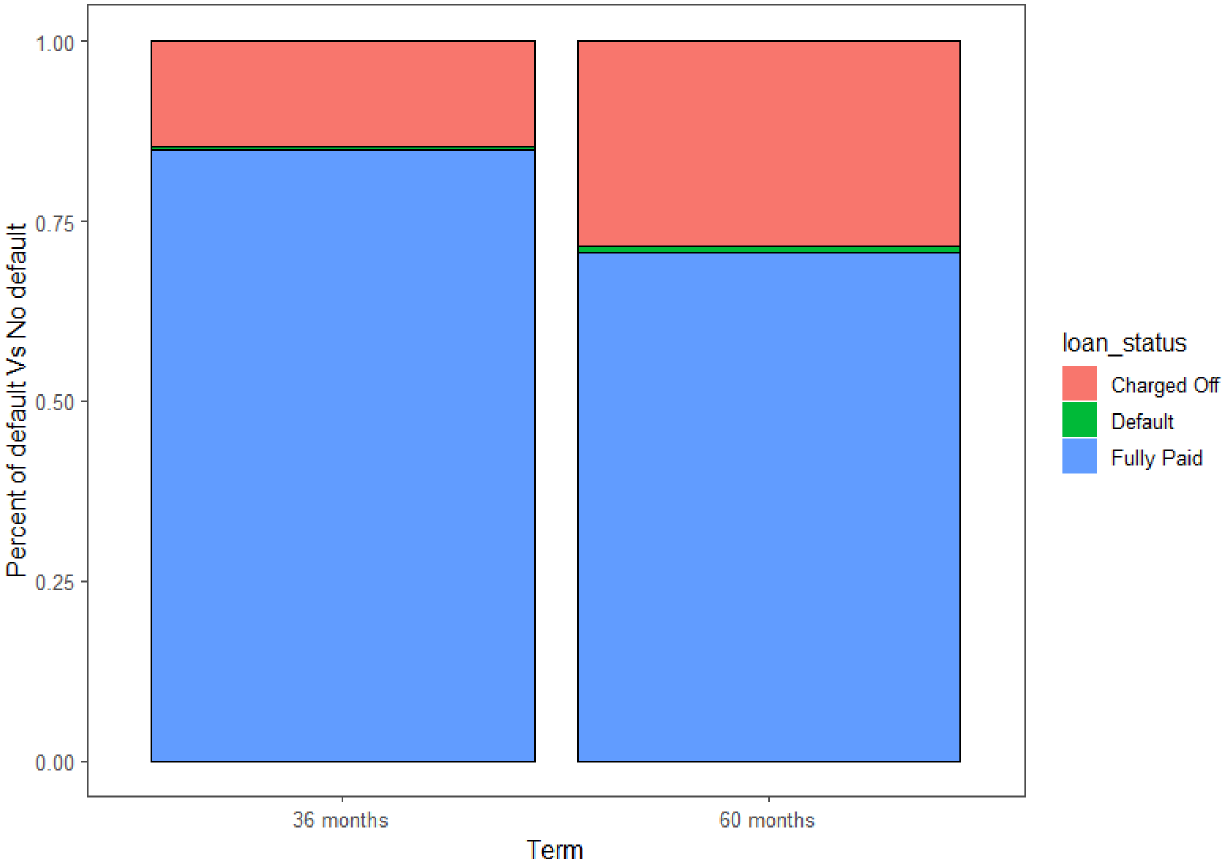


Figure 26: Loan status versus term

As we can see, it clearly seems to be a correlation between loan status and period of repaying. 60-month term loans seem to default or charged off more compared to 36-month loans. Lenders who prefer to repaying their loan in a shorten period are more likely to confront on their obligations compared to those who prefer a longer repaying period. The percentage of Bad Loans is 30% for 60-month term loans, almost double from the 18% for 36-month loans. Therefore, the explanatory variable "term" seems to be a good predictor for our classification models.

Next, we shall investigate the impact of grade on loan status. We believe that the better the grade, the smaller the probability of default, given that Lending Club's mechanisms on assessing loans' grades are working efficiently. In **(Figure 27),** loan status versus grade is plotted in a stack bar:

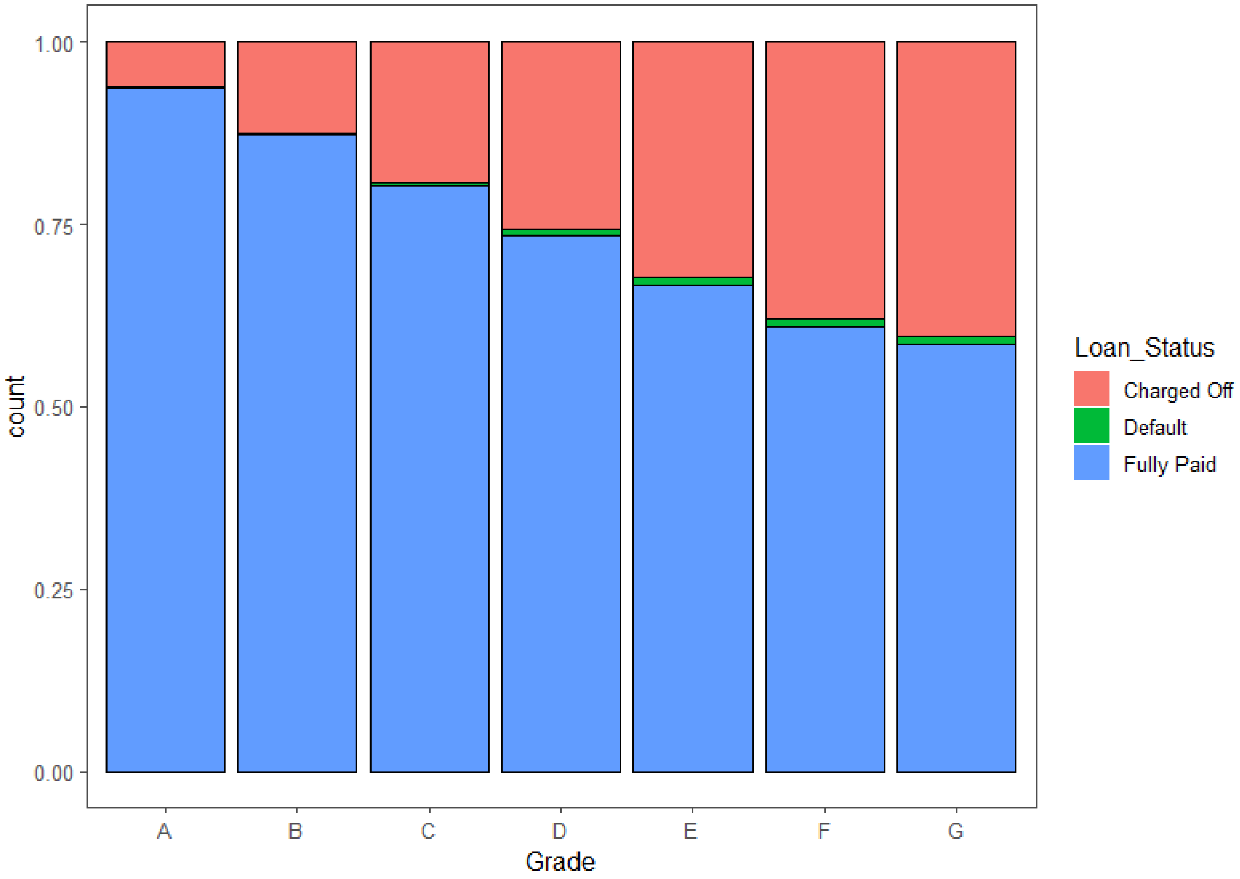


Figure 27: Loan status vs Grade

Stack bar reveals that indeed, there seems to be a relationship between loan status and grade. In fact, it seems that default increases with increase in Grade from A-G, A means lowest risk of loan default and G means higher risk of loan default. The better the grade of a loan, the smaller the probability to default or charged off.

Employment length is another variable we are going to investigate towards final loan status. Employment length is the prior occupational period of the applicant by the time they apply for the loan. We have already seen in chapter 7.2.1 that the majority of loans in our dataset were issued on applicants with more than 10 years of occupational experience. Now, we will compare loan status versus employment length in order to see if there is a certain group that heavily affects the probability of default **(Figure 28):**

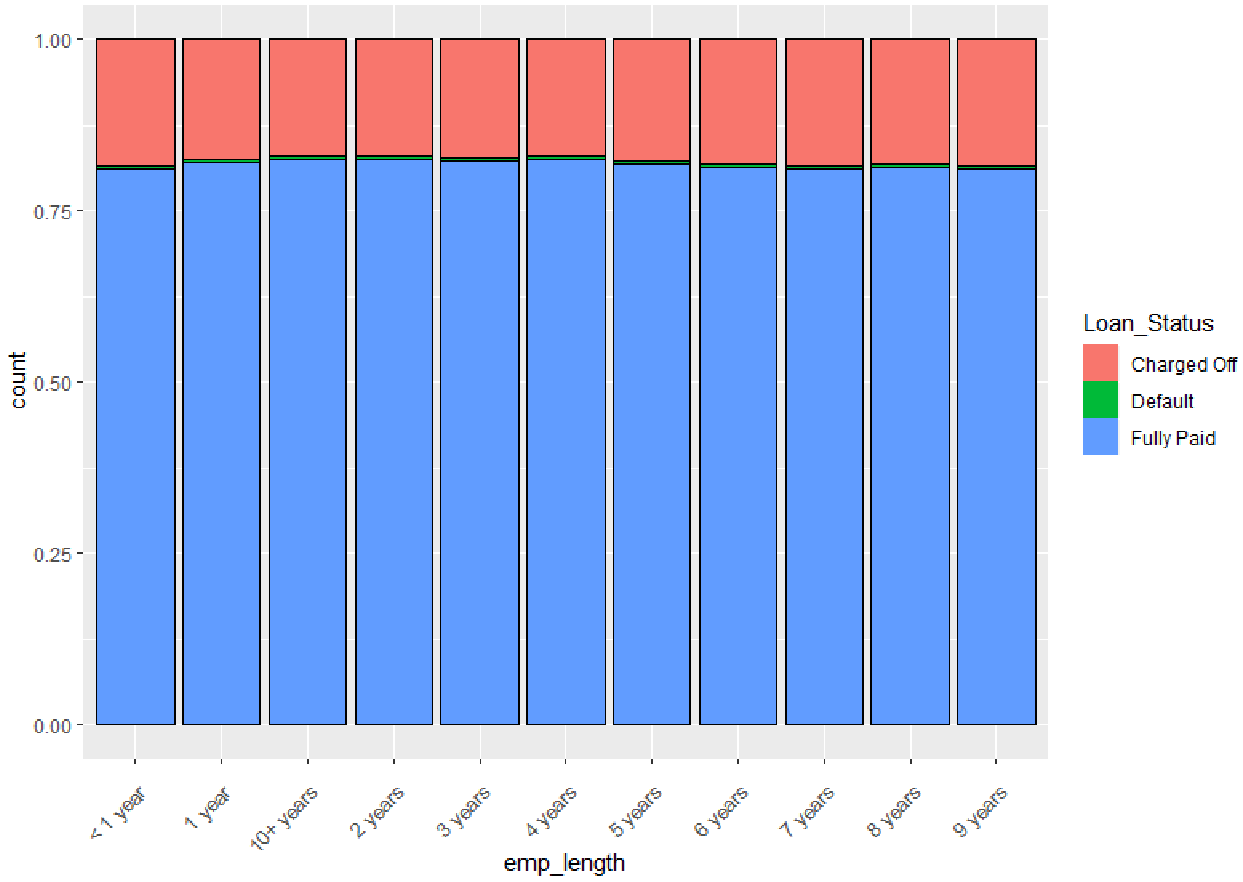


Figure 28: Loan status versus employment length

It looks like employment length does not solely drives default. There is no category that diverges compared to the others in terms of default. This finding does not mean that emp_length is a bad predictor for the assessment of the target variable as it may influence the classification algorithm in combination with other explanatory variables. This assumption will be put under testing during the modelling process. But for now, let us just keep in mind that there seems to be no correlation between *“emp_length”* and final loan status.

In **(Figure 29),** we have plotted loan status versus home ownership. Again, we come up with the conclusion that there is no significant difference between final loan status and type of home ownership. Applicants who rent a house have a slightly higher probability to default compared to the other categories, but in fact this difference is not significant compared to the corresponding probability for the other home ownership categories.

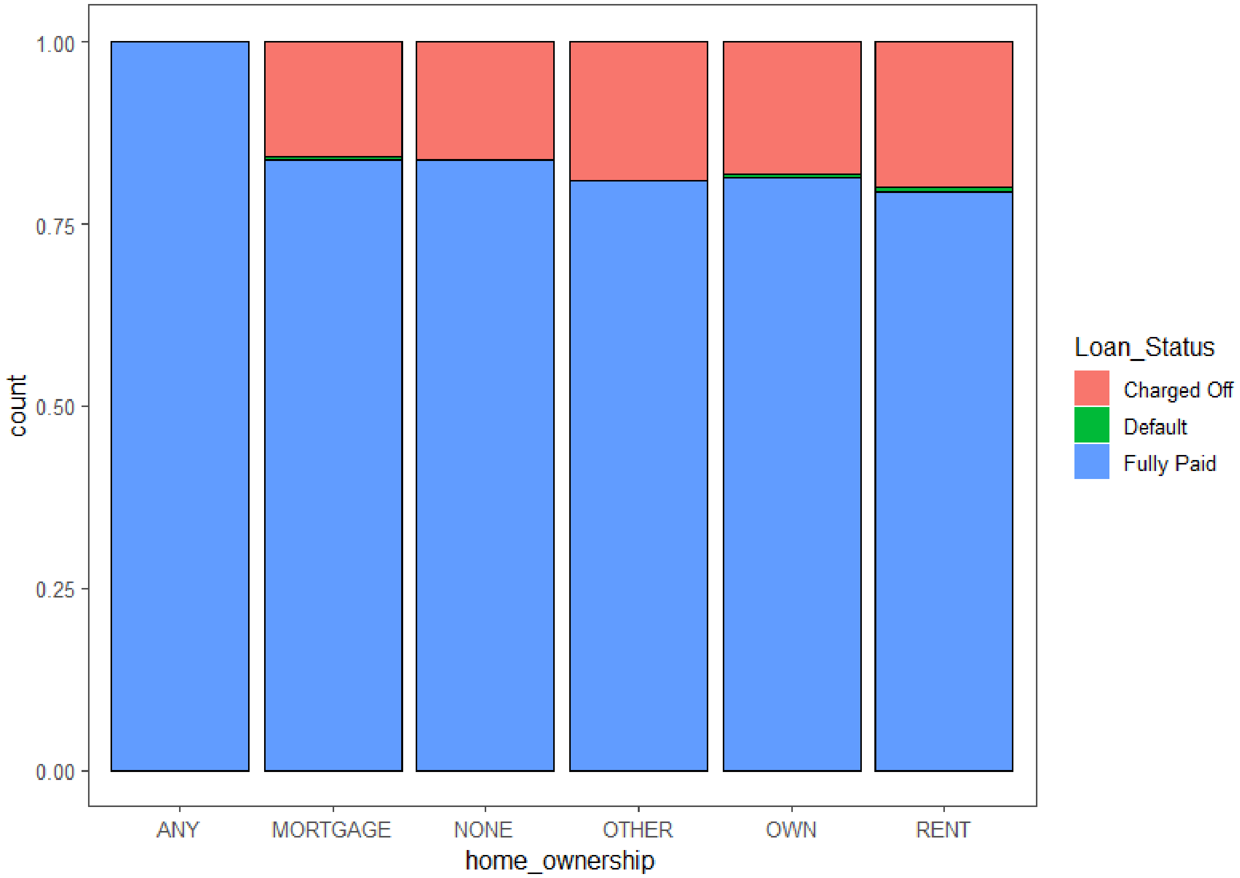


Figure 29: Loan status versus home ownership

The next explanatory variable we put under investigation is the loan purpose. Every applicant informs Lending Club for the purpose that he or she needs to be funded. Loan purpose is a categorical variable with 14 different categories. We have already described loan purpose and found that most of the applicants ask for a loan in order to consolidate a previous dept, with credit card repaying coming second in the reasons of application. In **(Figure 30),** we have plotted loan status versus loan purpose:

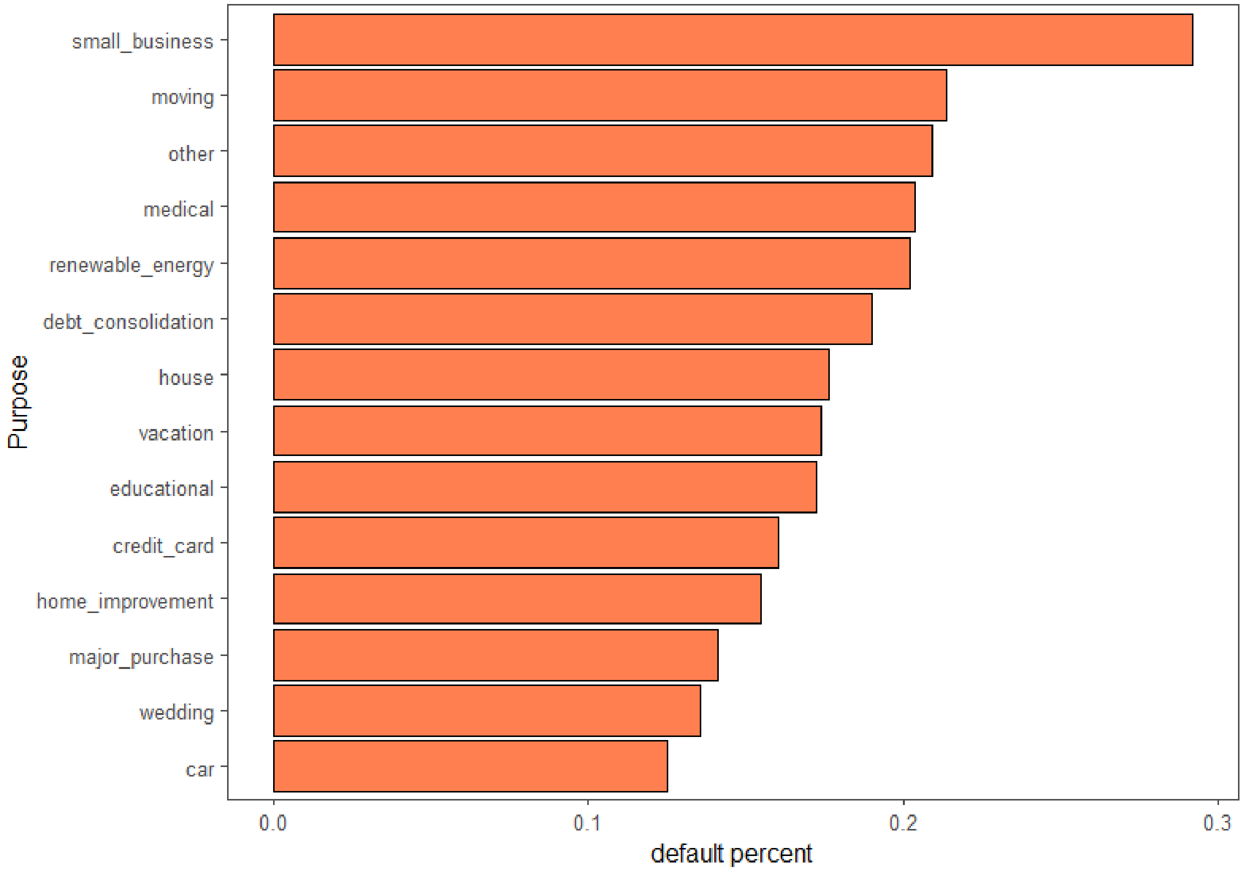


Figure 30: Loan status versus loan purpose

We can see that small businesses are more likely to default compared to the other categories with a cumulative probability of 30%. Dept consolidation and credit card seem to be less probable to default, both with a default probability less than 20%. Wedding funding and new car acquisition are the categories with the smallest probability to charge off.

Now, we will investigate how loan amount affects the final loan status. We assume that there should be a relationship between those variables and that loan amount strongly influences the probability of default. We have already explained that Lending Club policy indicates that the maximum amount of money an investor can borrow is 40000$. In **(Figure 31),** we can see the density plot of loan amount in terms of loan status:

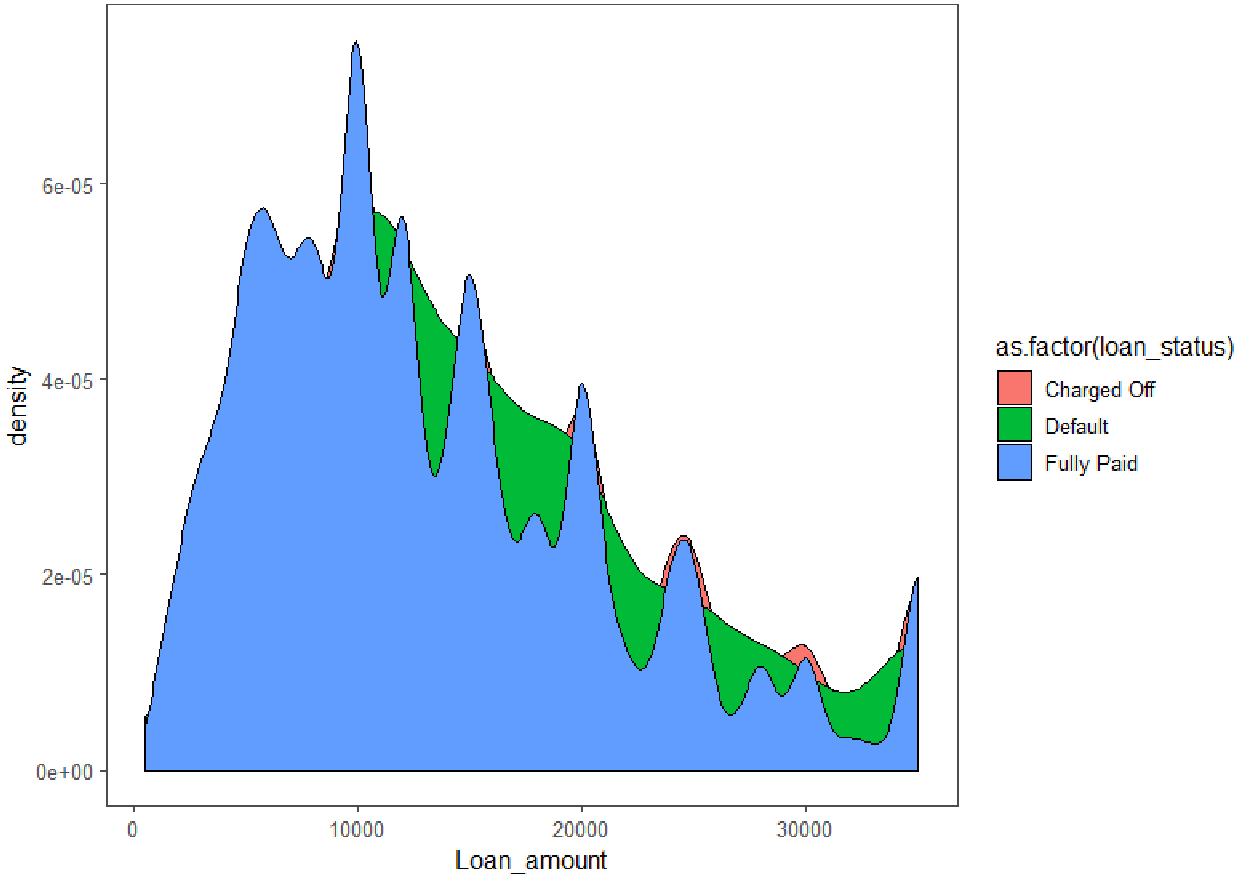


Figure 31: Density plot of loan amount in terms of loan status

There is obviously a clear relationship between the amount borrowed and the final status of the loan. As we can easily derive from the density plot, for amounts smaller than 10000$, the vast majority of loans are eventually fully paid by the borrowers. However, incidences of loan default can be seen when the loan amount is above 10,000$. As the amount funded increases, the probability of default becomes bigger. Loan amounts higher than 20000$ seem to have almost 50% probability to default. Therefore, our initial assumption that amount funded influences loan status seems to hold right and thus, “*loan_amount*” seems to be a good indicator for the prediction of default.

Interest rate is another explanatory variable that may give us valuable information for the modelling process. Interest rate is a predefined percentage on the initial capital which increases the final loan amount. In 7.2.1 we have grouped interest rate in three categories: "Low Rate" (for interest rate less than 10%), "Medium Rate" (10-20 prc) and "High rate" (more than 20%). We have seen that the majority of loans belong to the second category ("Medium Rate") with almost 70% of loans having an interest rate between 10 and 20 percent.

We assume that the higher the interest rate, the more probable is the applicant not to be compliant to their obligations. In **(Figure 32),** loan status versus ungrouped, initial values of interest rate is represented with the use of a boxplot:

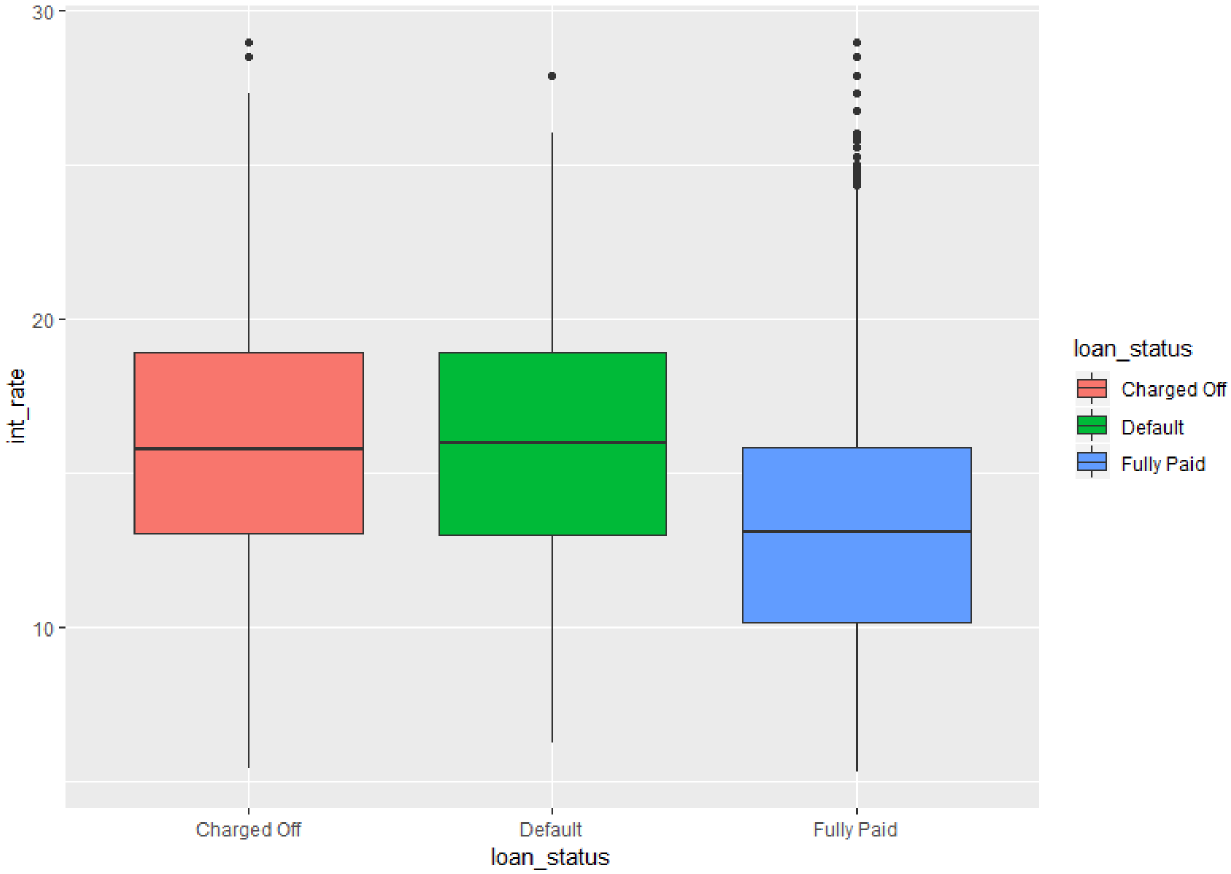


Figure 32: Boxplot of loan status vs interest rate

As we can see, loans finally characterized as “Bad” (Default and Charged off), seem to be driven by the predefined value of loan’s interest. The bigger the interest rate, the more probable the loan to default or charged off. In fact, the distribution of defaulted and charged off loans seems to be exactly the same if we look on the boxplot. They have the same median values (17%) and the same quartiles. The only difference is on maximum and minimum interest rates and in some outliers. On the other hand, “Fully Paid” loans generally have a smaller interest rate.

In general, we can say that higher interest rate is definitely linked to a greater number of defaults except for few outliers.

We are waiting for the same findings if we compare loan status versus the grouped variable that we have derived for interest rate, *“rate_category”*. We assume that “Medium Rate” and “High Rate” loans should demonstrate a higher probability to default or declared *“Charged Off”*. The corresponding plot is presented in **(Figure 33):**

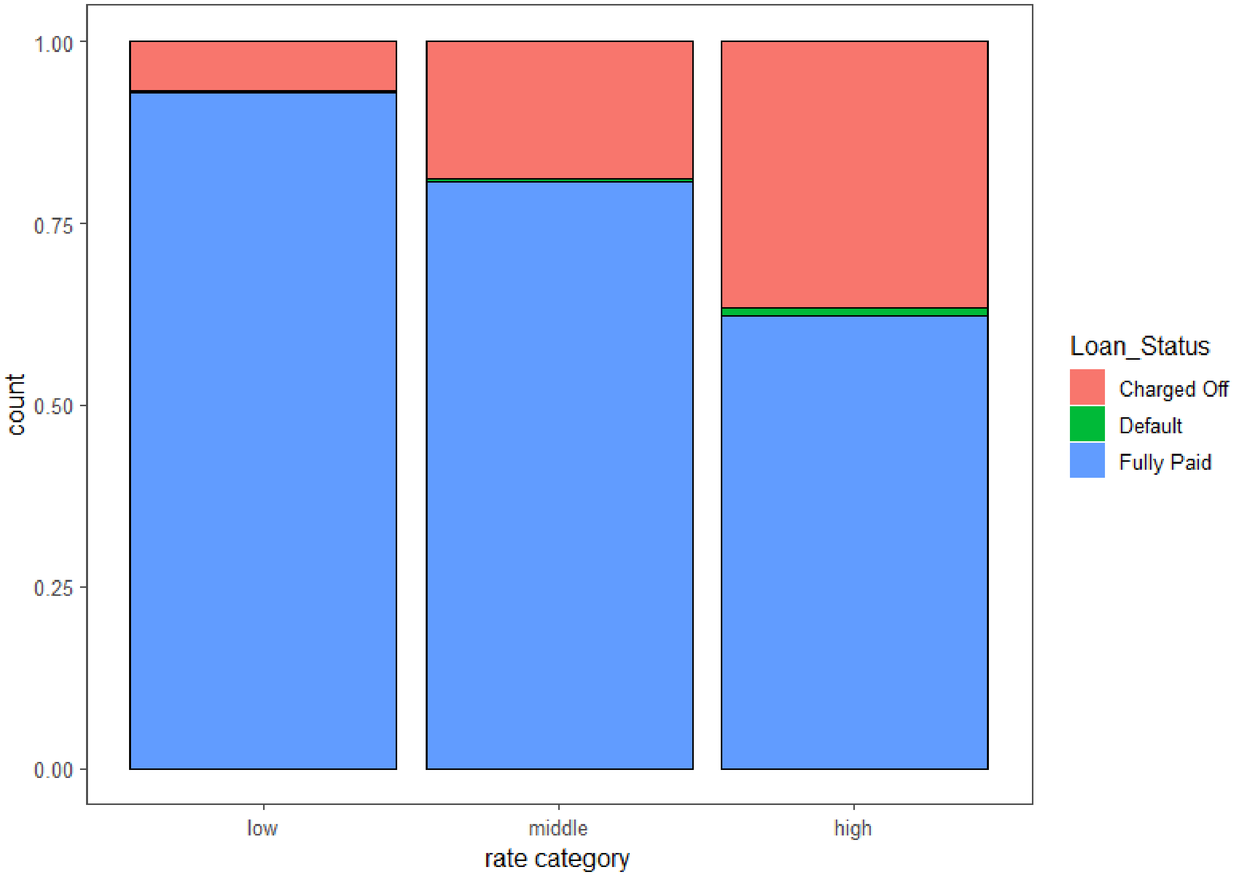


Figure 33: Rate category versus loan status

We can see that the relationship between the derived categorical variable *"rate_category"* and the target variable *"loan_status"* seems to be even stronger, compared to the raw initial variable. Almost 40% of "High Rate" loans tend to default or declared *"Charged Off"*. Loans with an interest rate of the first category ("Low Rate") tend to be "Fully Paid" by the borrowers in a vast percent of almost 90%.

Dept to Income (dti) is the ratio of the total amount borrowed to the annual income of the applicant. This variable implicitly describes the exposure of the applicant to the risk of default. We assume that the bigger the ratio of dept to income, the higher the probability for an applicant to eventually not be able to fulfill their obligations. Dept to Income versus loan ratio has been plotted in **(Figure 34):**

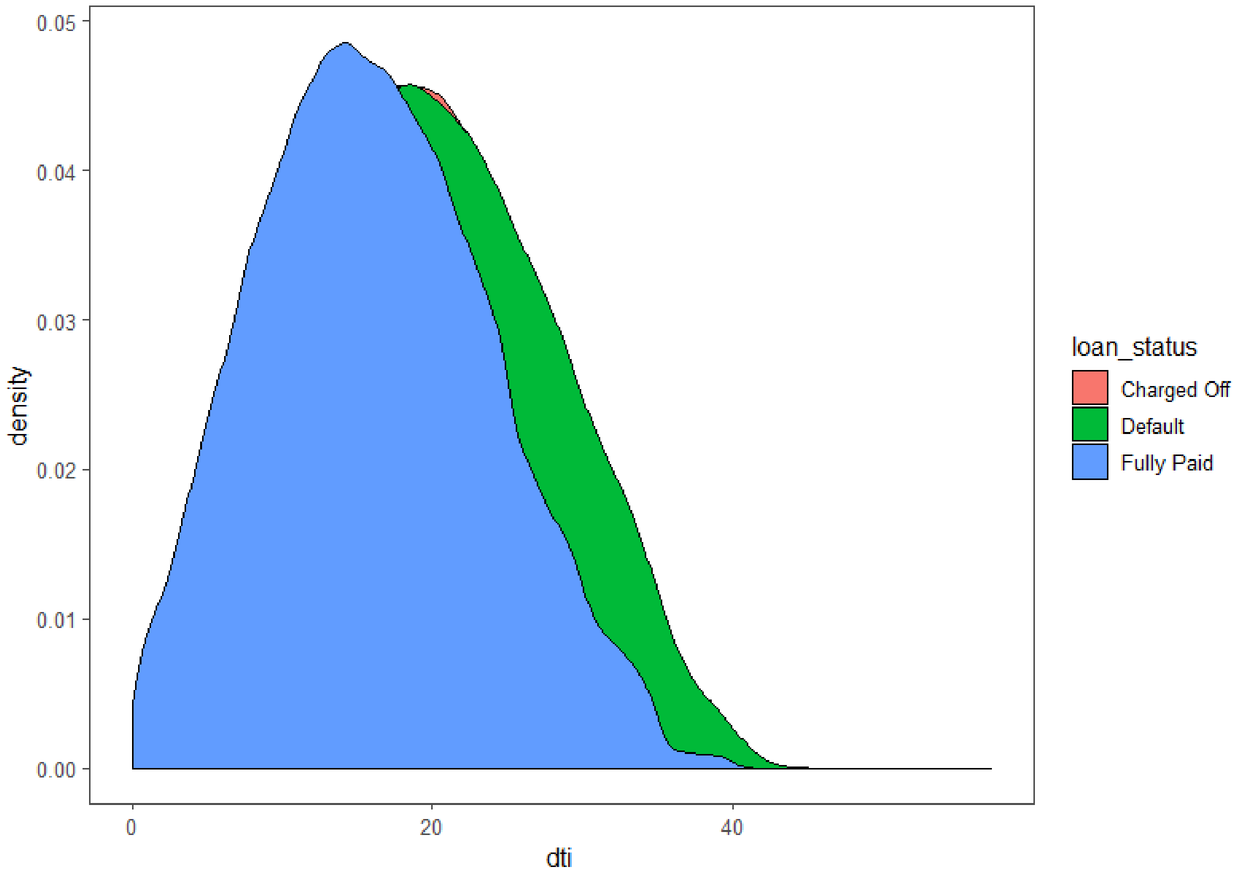


Figure 34: Dept to Income versus Loan Status

Form the density plot, we can easily understand how dti affects loan status. It seems that default increases when the dti is above 20; in other words, when the total dept obligations of an applicant are 20 times greater to his reported monthly income.

Annual income is another variable that may provide us with more insights towards the way it influences final loan status. Annual income is provided by lenders during their application for a loan. We have already investigated annual income on 7.2.1. We have found that the average annual income of the applicants is 75000$ and that 50% of lenders have an annual income of almost 65000$.

We will now check for possible relationship between this variable and loan status. For that, we have plotted the new, derived, grouped variable *"Income_Category" versus final loan status* **(Figure 35):**

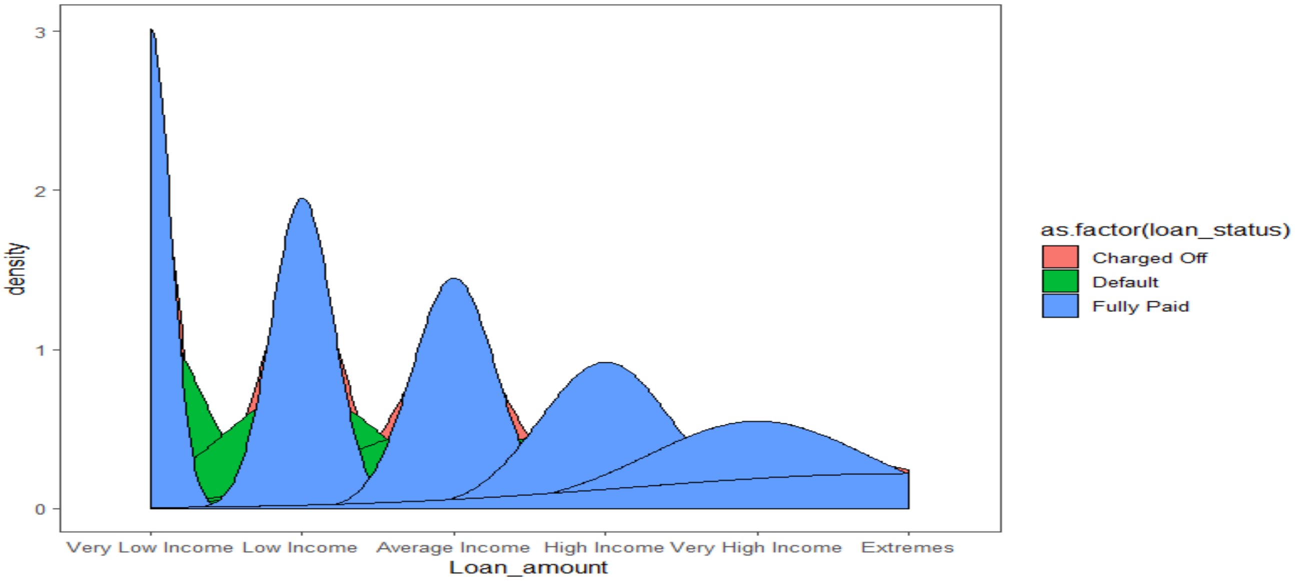


Figure 35: Income category versus Loan Status

Applicants who belong to the lowest layers of annual income have a greater probability to default compared to those with a higher annual income. The density of default loans is greater for Low

Income category and also the corresponding density of Charged Off category. Average Income applicants tend to be declared as "Charged Off" compared to the other five income categories.

Last but not least, and before stepping into the investigation of possible correlations between continuous explanatory variables, we will investigate a new derived metric, *"open_to_closed_accounts_ratio"* towards the final loan status of the applicants. *"open_to_closed_accounts_ratio"* is the ratio between all open credit lines of the borrower (credits that are in status current) and total credit lines an individual hold in their portfolio **(Figure 36):**

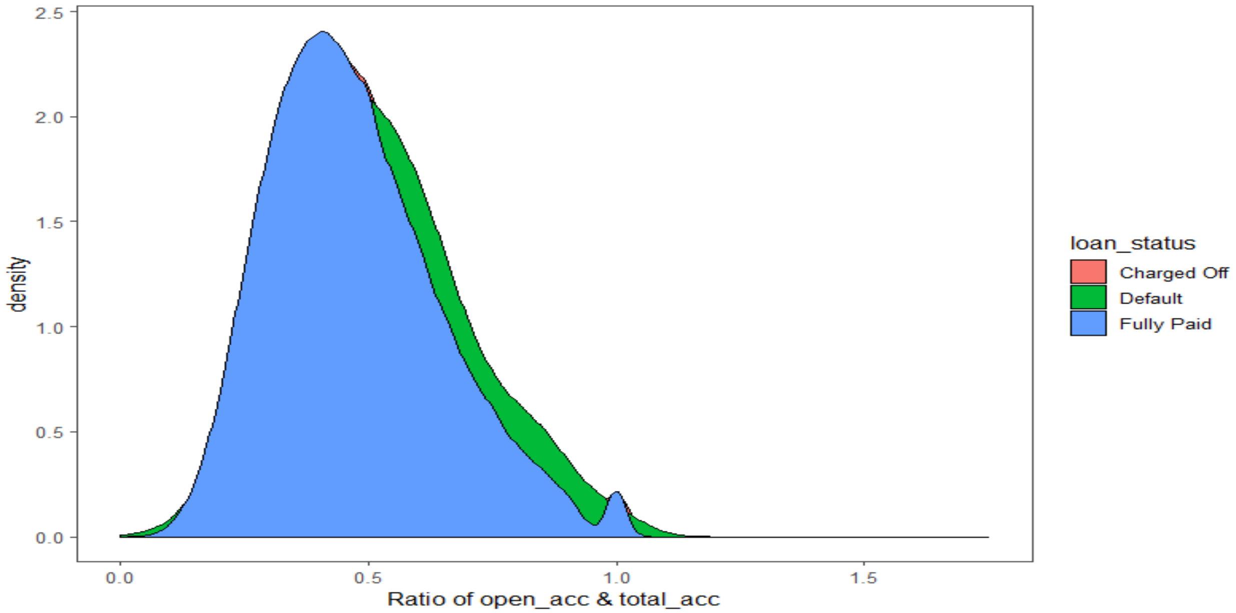


Figure 36: Open to close credit lines versus loan status

We can see that instances of default increases when the account ratio is above 0.5, in other words when open credit lines are more than half of total credit lines the borrower have on their portfolio when applying for a loan on the Lending Club. Therefore, we will take this new metric under review during the modelling process as it seems to be a good indicator for the prediction of default.

8.4.4 Correlations between continuous variables

A very important step in every data preparation process is to investigate all continuous variables in terms of correlations. In the broadest sense **correlation** is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a limited supply product and its price.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. However, in general, the presence of a correlation is not sufficient to infer the presence of a causal relationship.

We will search for correlations using Pearson Correlation coefficient. Pearson's correlation coefficient is a statistical measure of the strength of a linear relationship between paired data. In a sample it is denoted by r and takes values between -1 and +1. Minus 1 denotes an absolute negative linear correlation between two continuous variables when plus 1 denotes a corresponding positive linear correlation. If r equals zero, then the two variables are independent two each other (Ipsilantis, 2018).

For the calculation of Pearson Correlation Coefficient, we will use a built-in function provided in R statistical language. We have calculated the correlation coefficient for every pair of dataset's continuous variable and the result are demonstrated in **(Figure 37).** Correlation coefficient greater than 0.4 or less than -0.4 shall denote a strong positive or negative correlation between the corresponding variables. Otherwise, we shall argue that there is either weak or no correlation between the corresponding explanatory indicators. As Pearson correlation is calculated only for continuous variables, we have created a subset of our final dataset called *"loan_final_numeric"* which contains only the 12 numeric variables of our dataset plus the binary target variable we now have.

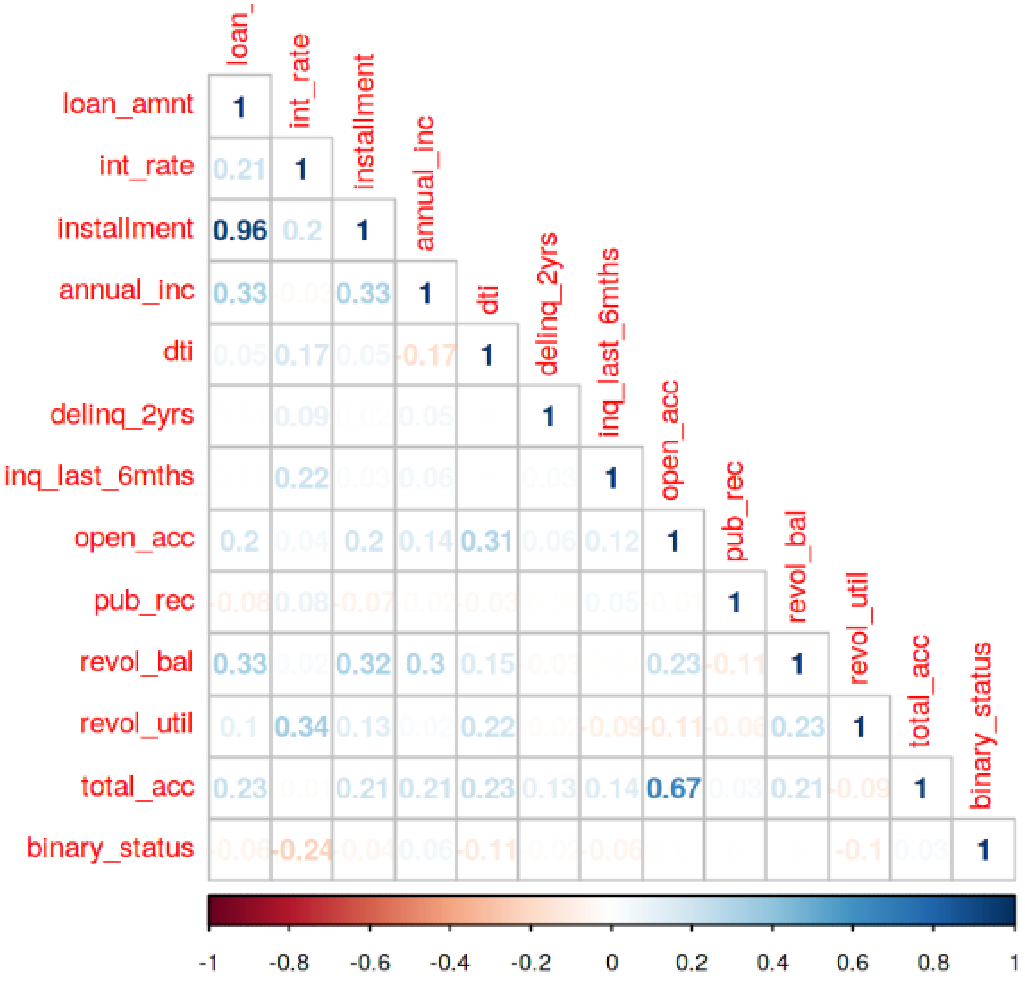


Figure 37: Pearson Correlation matrix

As we can see, there is a very strong positive correlation between *loan_amnt* and monthly *installment* (something obvious as the greater the amount funded the bigger the monthly installment) and also positive but weaker correlation between annual income and loan amount and installment. There is also a strong positive correlation between *total_acc* and *open_acc* variables. Additionally, weak negative correlation between our target variable and interest rate (r = -0.24) is observed and also the same between debt to income and annual income variables **(Figure 38):**

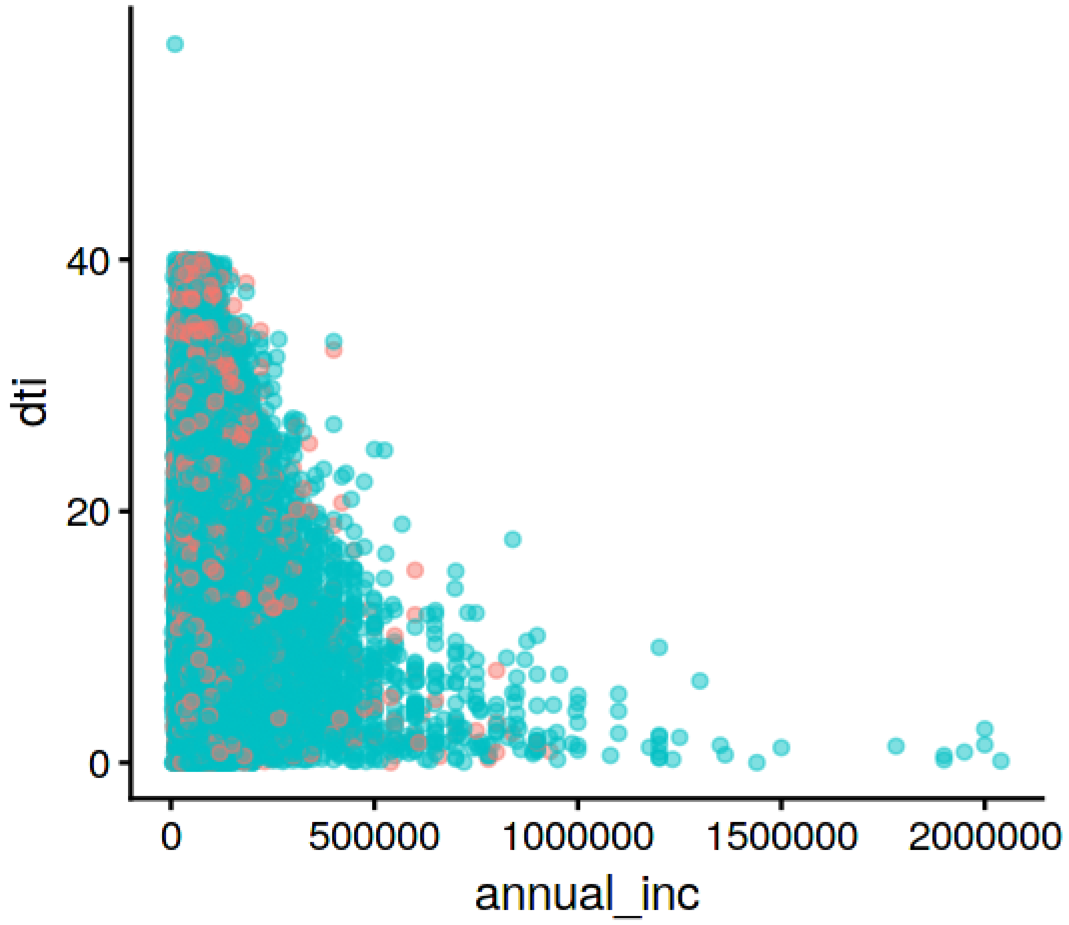


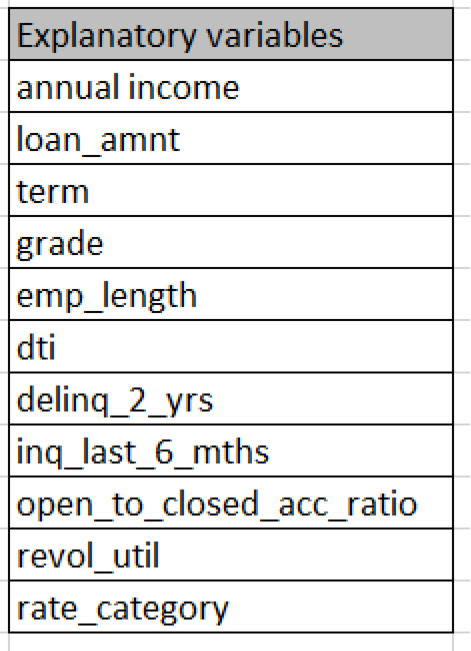
Figure 38: Negative correlation between dti and annual_income

The very strong positive correlation between loan amount and installment drives us to the decision to exclude installment variable from our analysis as it does not add incremental predictive power to our classification models given that loan amount will be used as explanatory variable. Open acc variable will not be initially excluded despite the strong correlation with total_acc variable but will be regarded as weak competitor in the list of input explanatory variables in ours models.

In that point, the data preparation process has been completed. We eventually concluded with a brand new dataset that included a total of 27 variables out of which 26 are explanatory ones and 254190 observations with loan status Fully Paid, which are denoted as “Good Loans” (value 1) and loan status “Default” or “Charged Off” which are characterized as “Bad Loans” (value 0). This dataset will be used as input data for our classification models. Before diving into the results of every single classification problem and compare their robustness and efficiency, a small description of the final input variable selection list will be introduced and also, we will argue towards the selection of the final input variables.

8.5 Input variables and imbalanced observations

Descriptive statistics are a very important step of every modelling process as it reveals significant information for the quality of the available variables. The analysis performed in the previous chapters in the initial dataset towards the decision of the final list of variables resulted in the use of 8 performance indicators which were used as explanatory variables in our models. The list of final input variables is shown in **(Table 9):**



Explanatory variables
annual income
loan_amnt
term
grade
emp_length
dti
delinq_2_yrs
inq_last_6_mths
open_to_closed_acc_ratio
revol_util
rate_category

Table 9: Final input variables

The selection of the final input vector has been made in two steps. First, conclusions were derived from the descriptive analyses process on the 26 independent variables and those findings were taken into account both in terms of statistical and predictive significance. Next, we constructed an initial logistic regrettor and asked for a threshold of less than 95% of statistical significance of a variable in order to discard this variable from the input variable's vector. Variables with score more than 95% percent are those that shown in **(Table 9).** Additionally, those 11 variables resulted in better relationship with the target variable as we have already described on chapter **(7.3.3).**

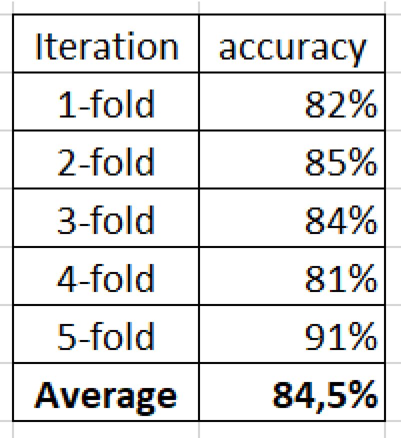
Another issue unveiled during the implementation of all selected models, was the problem of imbalanced observations. The percentage of Bad Loans into our final dataset is low compared to the majority class of Good Loans (18% of observations are characterized as "Defaulters" or "Charged off loans"). The issue of imbalance datasets is a very frequent phenomenon and was clearly revealed during the initial steps of our analysis. All 4 models were first trained on the initial, imbalanced training sets and the prediction results as well as the validation metrics were recorded. Next, we resampled the initial dataset set using the SMOTE resampling technique and repeated the modelling process. All four models performed significantly better both in terms of accuracy and sensitivity after SMOTE resampling. The results confirmed the bibliography towards the incremental predictive power of classification algorithm with dataset resampling on imbalanced observations.

Chapter 9
Results

In the final chapter of this Master Thesis, we introduce the results of every algorithm implemented before and after the resampling process and compare those algorithms in terms of sensitivity and accuracy. We also argue towards the multiple advantages for a financial institute to assess credit risk with the use of intellectual classification models.

9.1 Logistic Regression

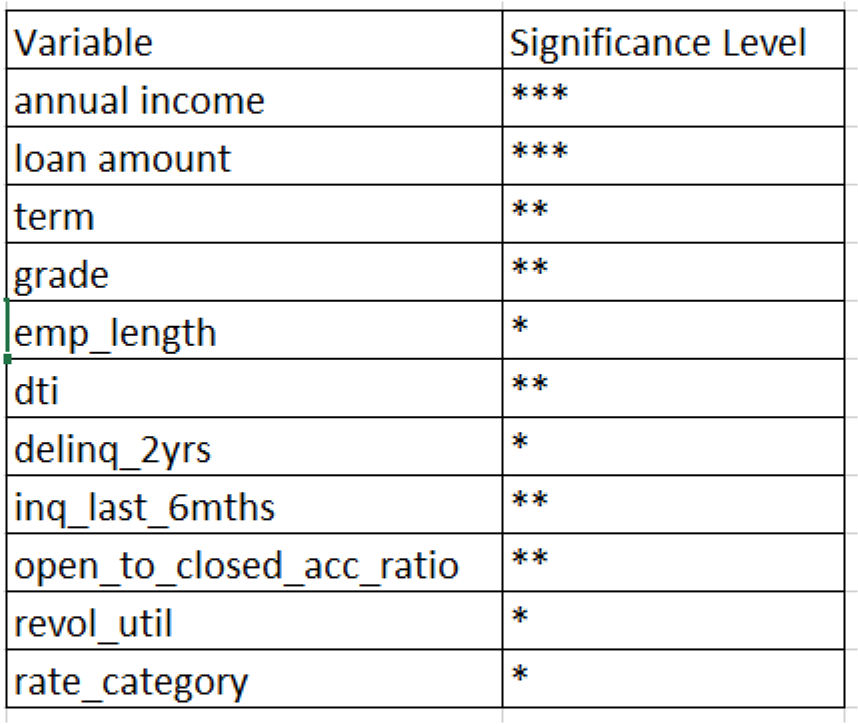
The first model tested for the prediction of Bad creditors is Logistic regression. We have implemented a logistic regression classifier using the built-in function *“glm”* provided in R. A 5-fold validation method implemented on the test set resulted in an average accuracy of 84.5%. The exact accuracy of every single iteration is demonstrated in **(Table 9):**



Iteration	accuracy
1-fold	82%
2-fold	85%
3-fold	84%
4-fold	81%
5-fold	91%
Average	**84,5%**

Table 10:5-fold validation results of Logistic regression

The statistical significance of each explanatory variable in a significance level of 5% is shown in **(Table 10):**

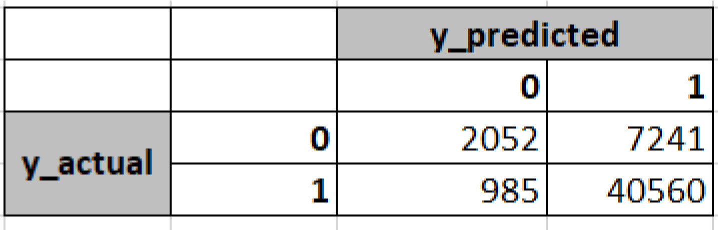


Variable	Significance Level
annual income	***
loan amount	***
term	**
grade	**
emp_length	*
dti	**
delinq_2yrs	*
inq_last_6mths	**
open_to_closed_acc_ratio	**
revol_util	*
rate_category	*

Table 11:Significance level of explanatory variables

Higher significance variables are denoted with 3 stars (***), when lower significance variables are denoted with one star (*)

The false positive error (predictive result “Good Loan” when actual result was “Bad Loan”) were 7241, almost 75% of total actual “Bad Loans”. On the other hand, true positive observations were 985 which means that 8% of Good Loans were characterized as possible defaults. Those results are shown on (Table 9):



		y_predicted	
		0	1
y_actual	0	2052	7241
	1	985	40560

Table 12: Confusion Matrix of Logistic Regression model

A threshold of 50% (cut off point) was used for the characterization of the predicted results into the corresponding categories. In other words, each predicted observation with probability higher than 50% was attached to the category "Fully Paid" when probability less than 50% denoted a "Bad Loan" prediction. The threshold is arbitrary and was decided in such a high level in order to minimize as possible the false positive results of the predictor.

Area Under Curve (AUC) and Receiving Operator Characteristics Curve (ROC) were also calculated for the assessment of the accuracy and predictive power of the Logistic Regression model **(Figure 38):**

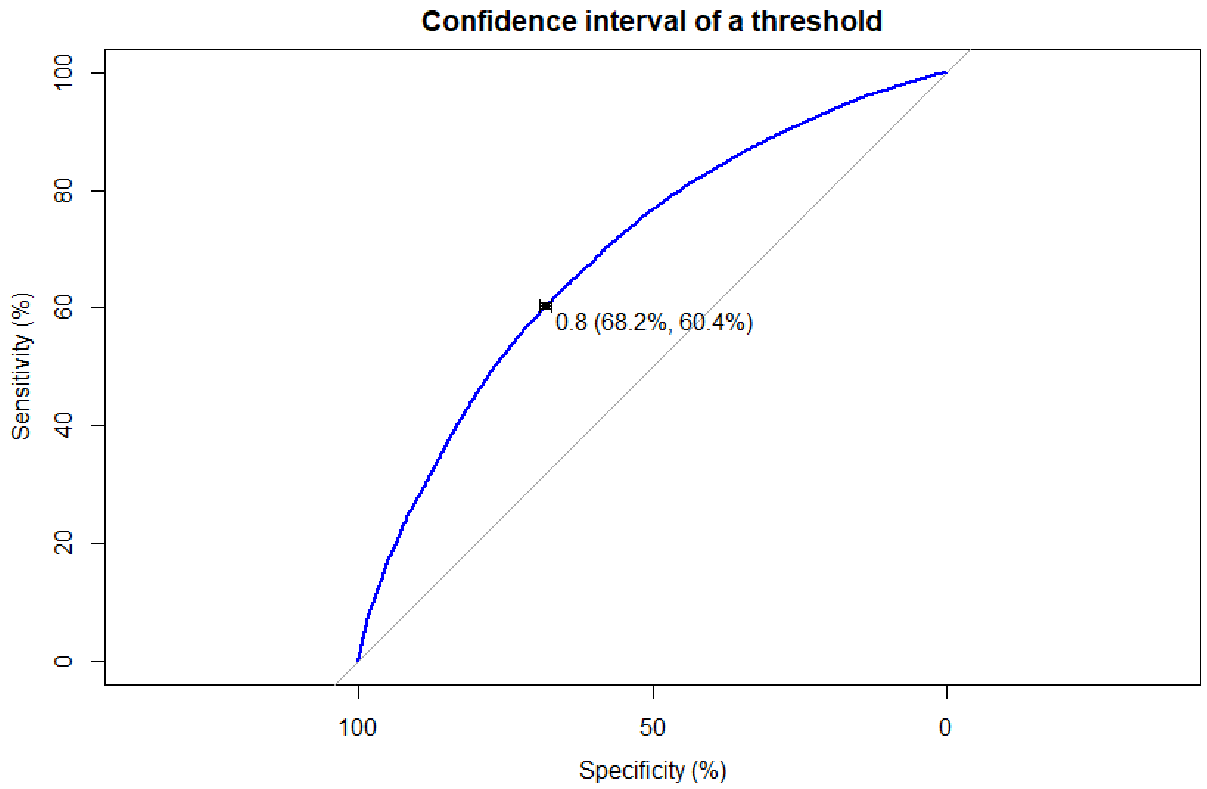
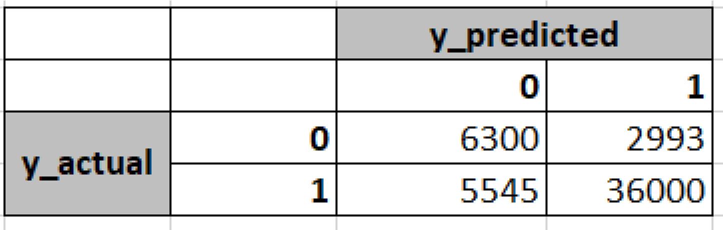


Figure 39: AUC and ROC curve of Logistic Regression classifier

The accuracy of the model is almost 3% above the actual percentage of “Bad Loans” in our final dataset. True positive percentage is 97% which means that logistic regression model predictes accurate the actual “Fully Paid”. The percentage of recall however was also very small; only 28% of actual defaulters were predicted by logistic regression model. Therefore, this model also resulted in the initial problem of data imbalance.

We proceed in resampling the initial dataset using SMOTE sampling. After resampling, the logistic regrettor was recalculated and resulted in a better prediction of the not dominant class of the target variable. The k-fold cross validation revealed a better overall average accuracy (89%), but with significantly better percentage of True Negative predictions:



		y_predicted	
		0	1
y_actual	0	6300	2993
	1	5545	36000

The percentage of True Positives has slightly increased (86%), but the important outcome is that recall percentage is almost 70%, indicating a significant better prediction of the true defaulters. Area under curve has also significantly increased, from 69% to 82%.

The ROC curve was also recalculated over the new model. It showed a significant improvement on the sensitivity and predictability power of the model driven by the significant increase on the True Negative predictions' percentage **(Figure 40):**

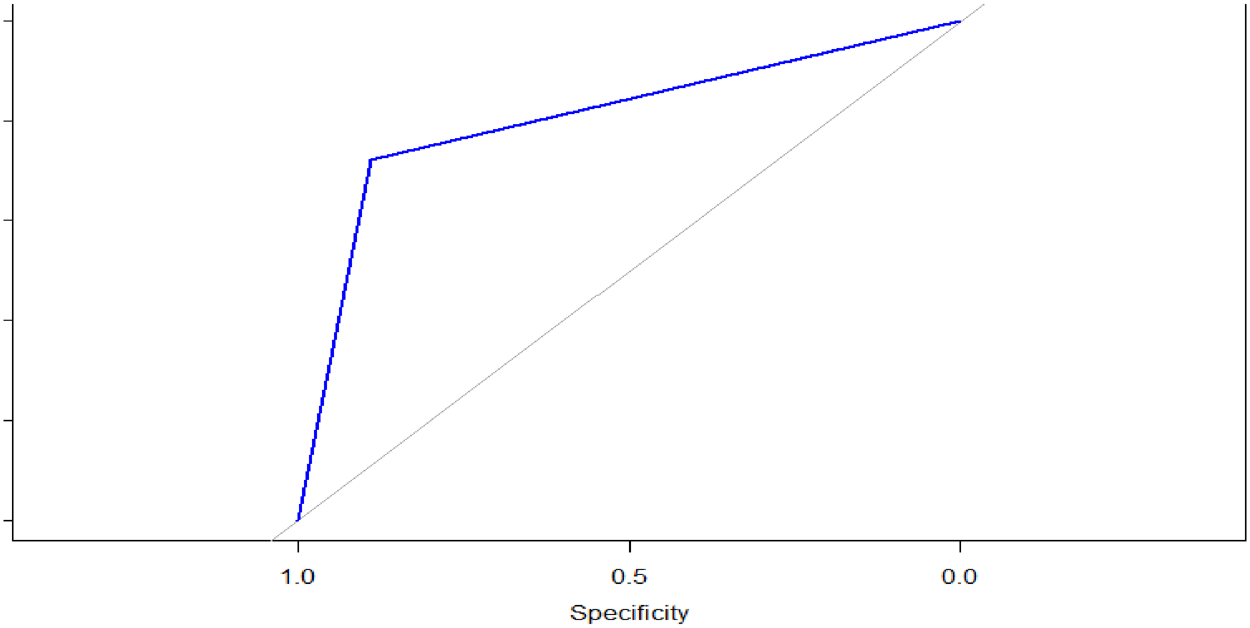
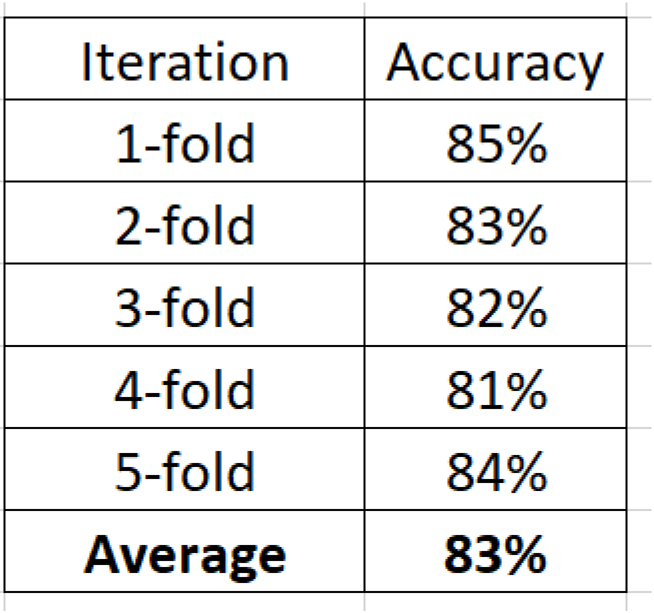


Figure 40: ROC Curve of Logistic Regression classifier after balancing the initial observations.

9.2 Classification Trees

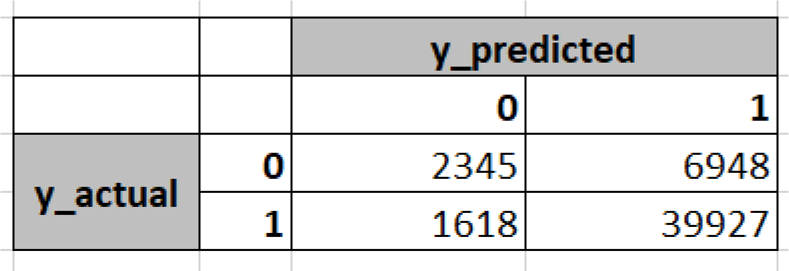
The second model implemented and investigated towards the prediction of actual defaulters is Decision Trees. The implementation was made using the *“rpart”* library available in R Statistical Language. The initial dataset was split in 5 different training sets using 5-fold cross validation method, the same used to the Logistic Regression model. The accuracy of the 5 iterations as well as the average accuracy of the model trained in the initial dataset is presented in **(Table 13):**



Iteration	Accuracy
1-fold	85%
2-fold	83%
3-fold	82%
4-fold	81%
5-fold	84%
Average	**83%**

Table 13: 5-fold validation of Decision Tree model on the initial dataset

The average accuracy was 83%, driven from the high percentage of True Positive prediction (actual Fully Paid loans). The confusion matrix in **(Table 14),** illustrates the initial results of the model:

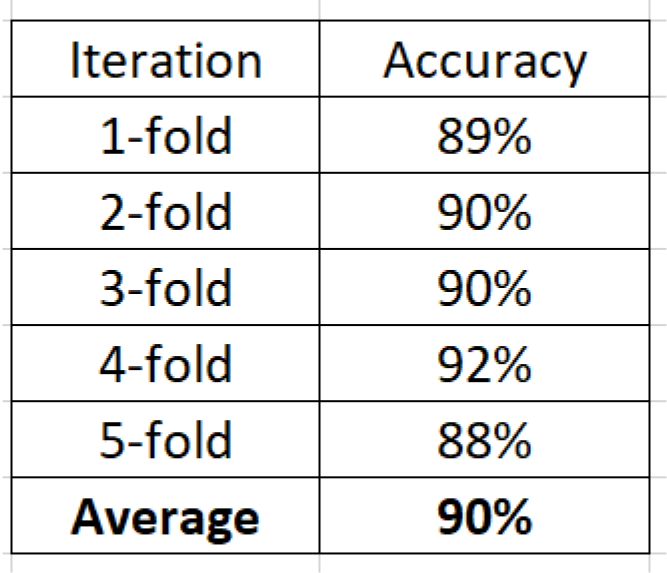


		y_predicted	
		0	1
y_actual	0	2345	6948
	1	1618	39927

Table 14: Confusion Matrix of the test set with Decision Trees initial imbalanced dataset

The True Negative percentage is slightly better than the corresponding of the Logistic Regrettor (25%) but far from accepted in terms of predicting the actual defaulters. Area Under the Curve was 0.70. The percentage of True Positive prediction is in the same level with Logistic Regression.

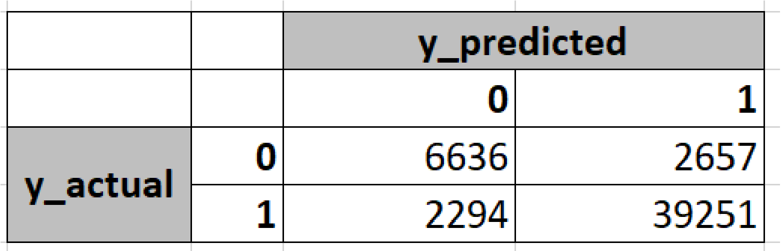
SMOTE resampling on the initial training set was performed and again the model was trained in the new dataset. The 5-folds' accuracy after the resampling is demonstrated in **(Table 15):**



Iteration	Accuracy
1-fold	89%
2-fold	90%
3-fold	90%
4-fold	92%
5-fold	88%
Average	**90%**

Table 15: Accuracy of the 5 iterations after SMOTE resampling

The accuracy of the model has been significantly improved after SMOTE resampling as the synthetic observations of the minority class constructed aid the algorithm towards a better recognition and classification of the observations in the proper category. The sensitivity of the model has also improved and reached 72%. That means, the model can distinguish between the two classes of the target variable correctly with probability 72% **(Table 16):**



		y_predicted	
		0	1
y_actual	0	6636	2657
	1	2294	39251

Table 16: Confusion Matrix of the test set observations after resampling

Area Under Curve has also adjusted to 0.875 and the ROC curve demonstrates the better results of the model after SMOTE resampling **(Figure 41):**

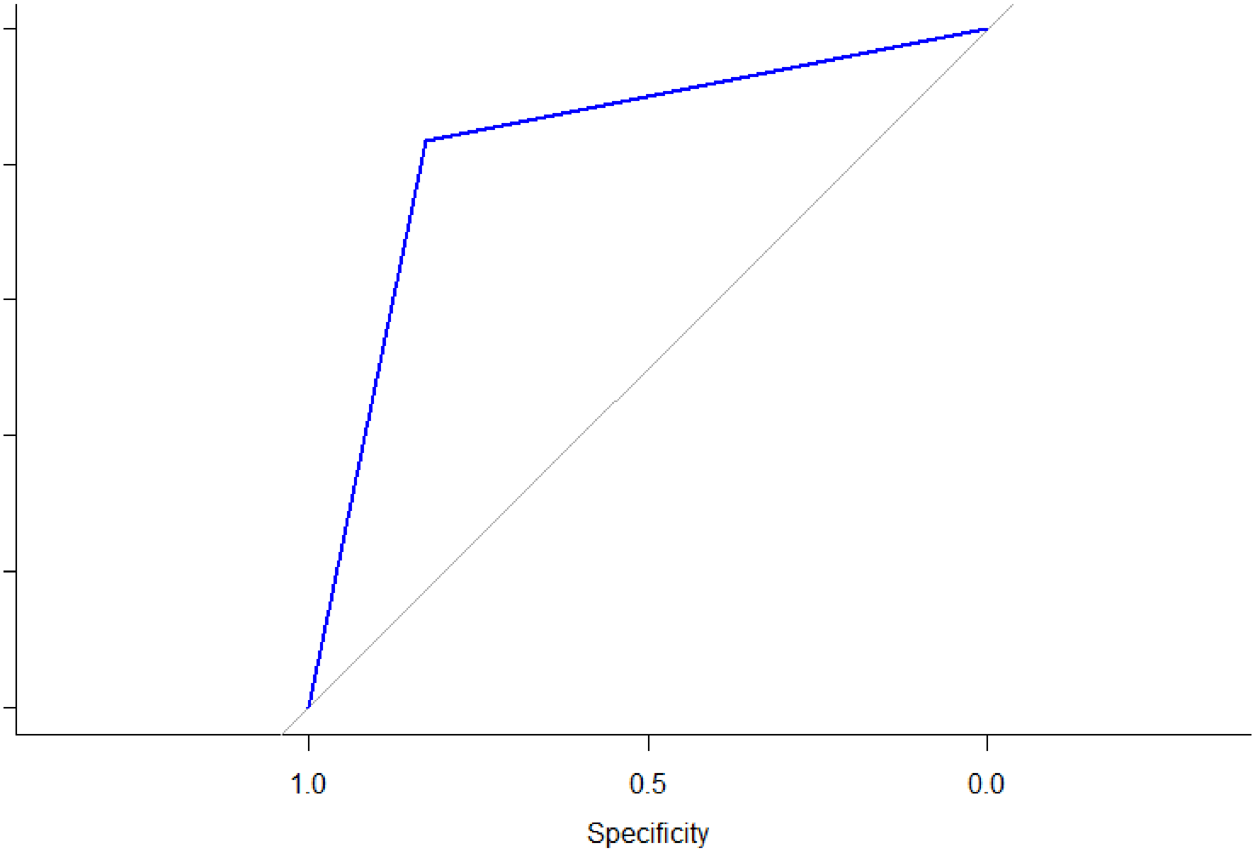
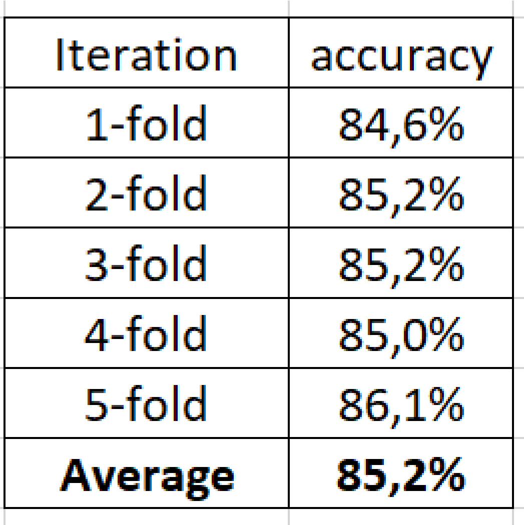


Figure 41: ROC Curve of Decision Tree Model after SMOTE method implementation

9.3 K-nearest neighbors

The last model implemented in this Master Thesis for the prediction of defaulters is k-NN neighbors. As already described in the theoretical framework of this thesis, K-NN is a nonlinear classifier, which is commonly used for segmentation models but also, has great predictability power as a classification algorithm. K-NN comes as a built-in function in R, into "*caTools*" library.

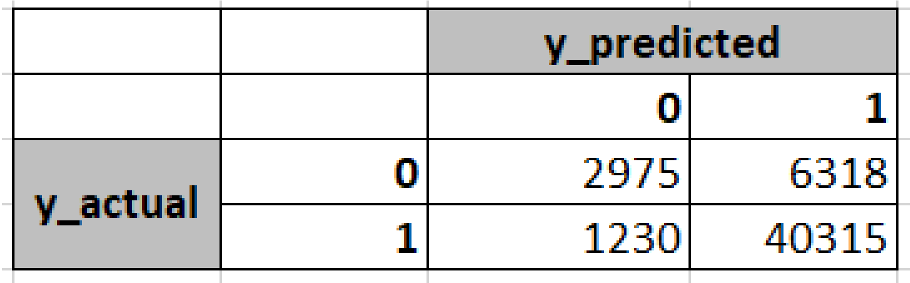
The model was trained on the initial, imbalanced dataset using the same explanatory variables with all 3 previous models. The 5-fold validation process resulted in an average accuracy of 85.2%. The accuracy of the 5-folds of the training set split are demonstrated in **(Figure 40):**



Iteration	accuracy
1-fold	84,6%
2-fold	85,2%
3-fold	85,2%
4-fold	85,0%
5-fold	86,1%
Average	**85,2%**

Figure 42: 5-fold cross validation of K-NN model

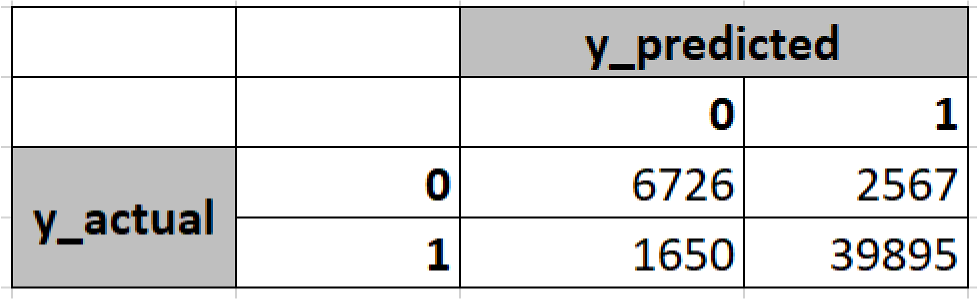
True positive percentage is 96% but – as with the other models- the percentage of True Negative predictions is low (32%) **(Table 13):**



		y_predicted	
		0	1
y_actual	0	2975	6318
	1	1230	40315

Table 17: Confusion Matrix of K-NN classifier before SMOTE resampling

The model run again after balancing the initial dataset. The method used for manipulate the imbalanced observation was SMOTE sampling. The 5-fold validation method unveiled a much better average accuracy of 91%. The percentage of True Negatives has been significantly improved, reaching 74% of Bad Loans' correct prediction. The Area Under the Curve has also increased at 85%. The ROC curve and the confusion matrix of the K-NN algorithm on the balanced training set are represented below **(Table 14), (Figure 41):**



		y_predicted	
		0	1
y_actual	0	6726	2567
	1	1650	39895

Table 18: Confusion Matrix of K-NN classifier after resampling

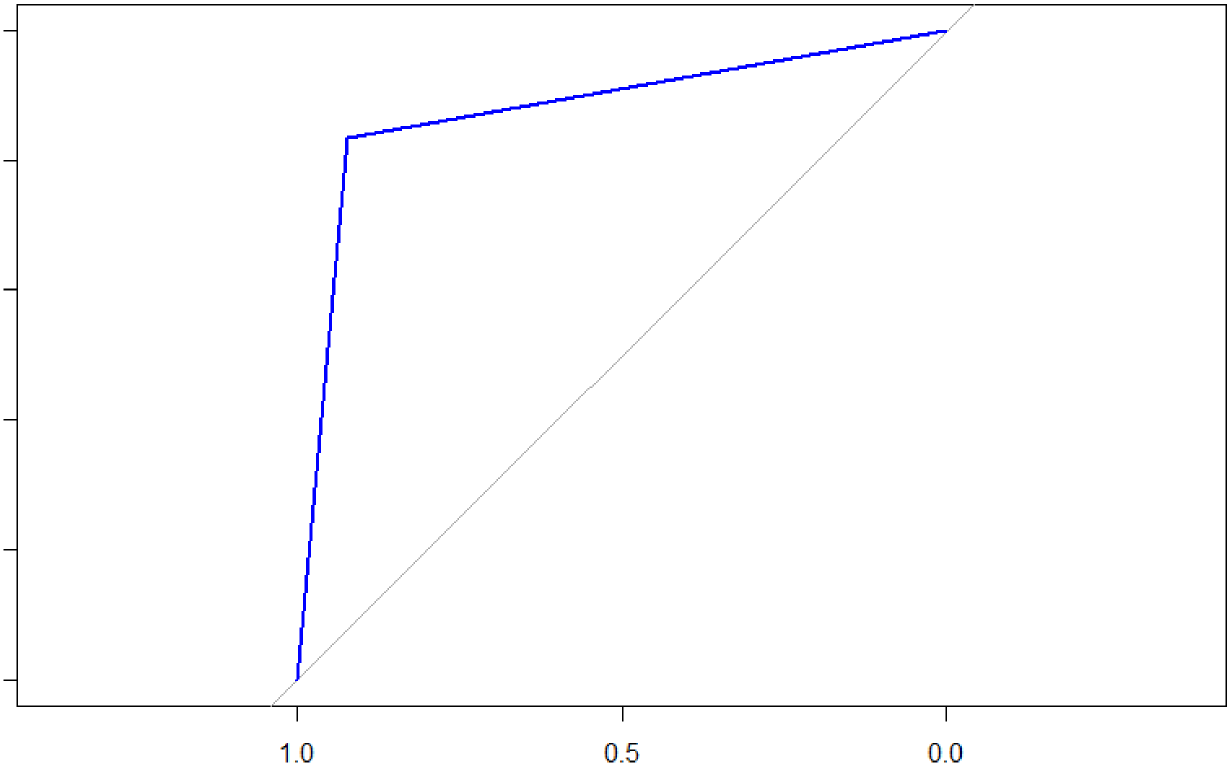
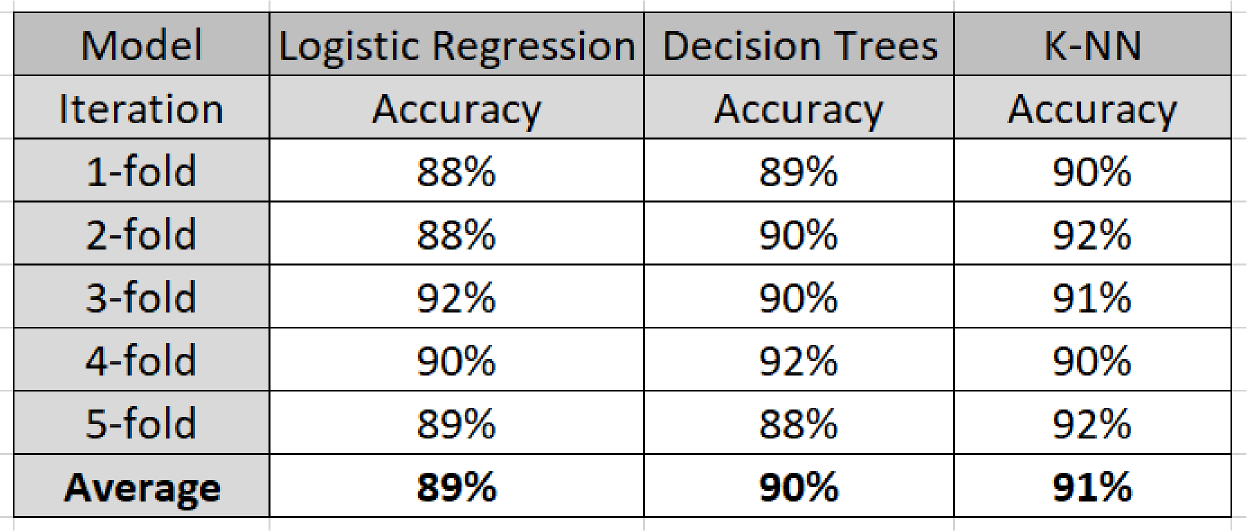


Figure 43: ROC Curve of K-NN classifier after balancing the initial observations.

9.4 Discussion and Model Comparison

We have tested three different models in terms of their predictability in recognizing and correctly classifying instances of the target variable *"loan_amount"* into the correct class between Good and Bad Loans. Confusion Matrix, Area Under Curve and ROC curve are the evaluation methods used in order to recognize the model that best fits our analysis. The common issue between all 3 implemented models was the problem of imbalanced observations. The minority class "Bad Loans" had less than 20% of observations in the initial dataset, resulted in low performance of the models in terms of correctly recognizing defaulters and charged off loans (Bad Loans). We manipulated the imbalance issue with SMOTE resampling on the initial dataset. 5-fold cross validation method used for incremental stability and validation of models' results. Between all 3 classification algorithms put under comparison, K Nearest Neighbor classifier, seems to over perform Logistic Regression and Decision Trees towards the prediction of actual defaulters. Additionally, K-NN performed better also in terms of total model accuracy both before and after the resampling process **(Table 15):**

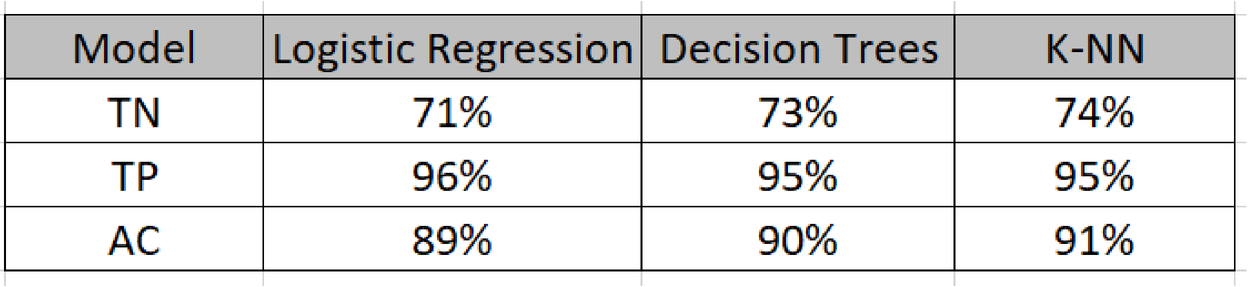


Model	Logistic Regression	Decision Trees	K-NN
Iteration	Accuracy	Accuracy	Accuracy
1-fold	88%	89%	90%
2-fold	88%	90%	92%
3-fold	92%	90%	91%
4-fold	90%	92%	90%
5-fold	89%	88%	92%
Average	**89%**	**90%**	**91%**

Table 19: Accuracy Comparison after SMOTE resampling

K-NN performed better in correctly recognizing minority class observations (True Negative proportion). In a business manner, sensitivity is of higher importance than accuracy as long as risk exposure of an institute relies on the percentage of possible defaulters rather than correctly recognizing profit opportunity depicting from a “Good applicant”. In other words, even if a model over performs another in terms of better True Positive percentage (as for example Logistic Regression does), the incremental value for a business should arise from minimizing the risk exposure towards possible depression. For different business sectors, the acceptance threshold for issuing a loan may be different and according to the overall strategy and decision making. In our case, we used a neutral threshold of 50% - that means credit score was calculated given that a Bad applicant was recognized by the algorithm if the predicted probability of default was higher than 50%.

A different threshold would suggest different results. For example, a threshold of 60% would give even better prediction of the minority class observations (better True Negative percentage) but with a lower proportion of True Positive recognitions (actual good applicants). This is the point where strategy and goal setting in the highest management level contributes to the assumptions made before implementing a classification model. **(Table 16)** represents final sensitivity and accuracy comparison of the 3 models after resampling the initial data:



Model	Logistic Regression	Decision Trees	K-NN
TN	71%	73%	74%
TP	96%	95%	95%
AC	89%	90%	91%

Table 20: Comparison Metrics (TP, TN, AC) of the 3 models

Logistic Regression slightly over performed Decision Trees and K Nearest Neighbors in terms of accuracy, but this difference is driven by the highest proportion of True Positive observations. Confusion Matrixes and comparison plot of ROC curves present a comparison view of the 3 implemented algorithms **(Figure 43):**

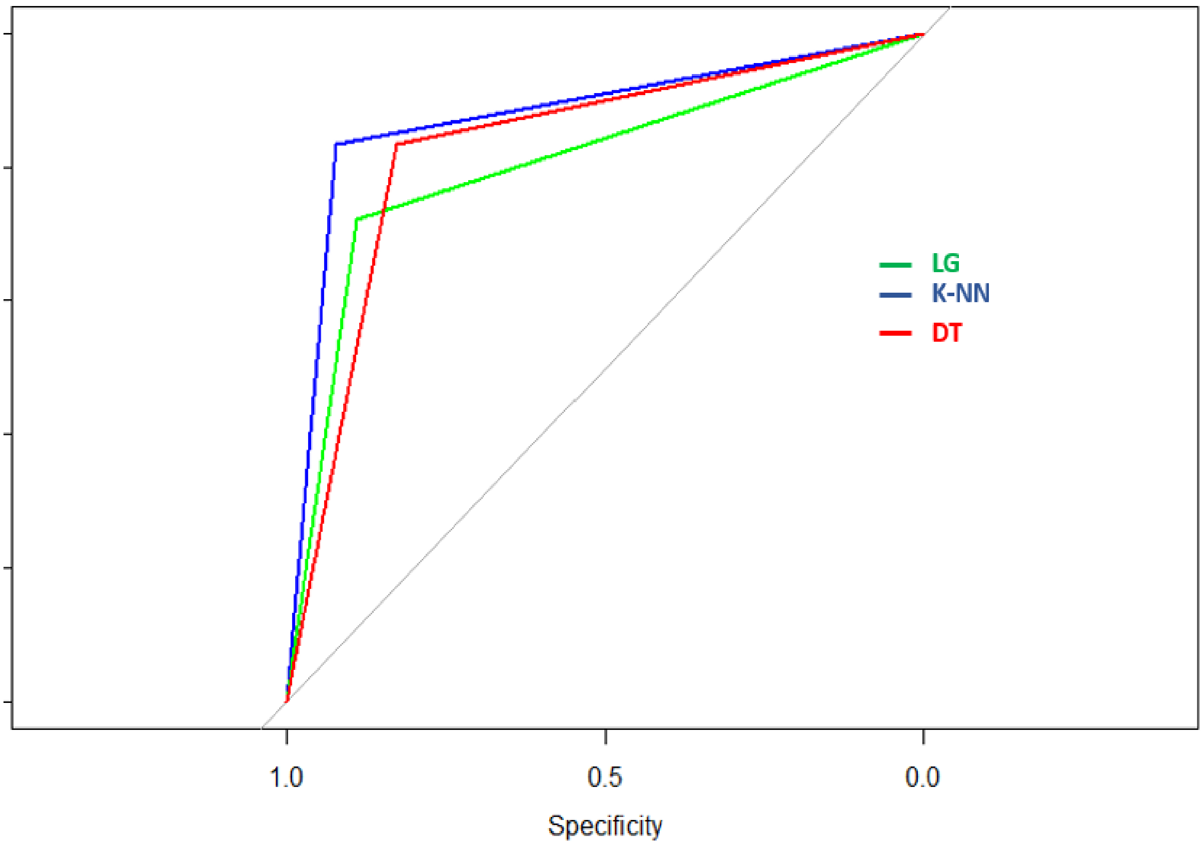
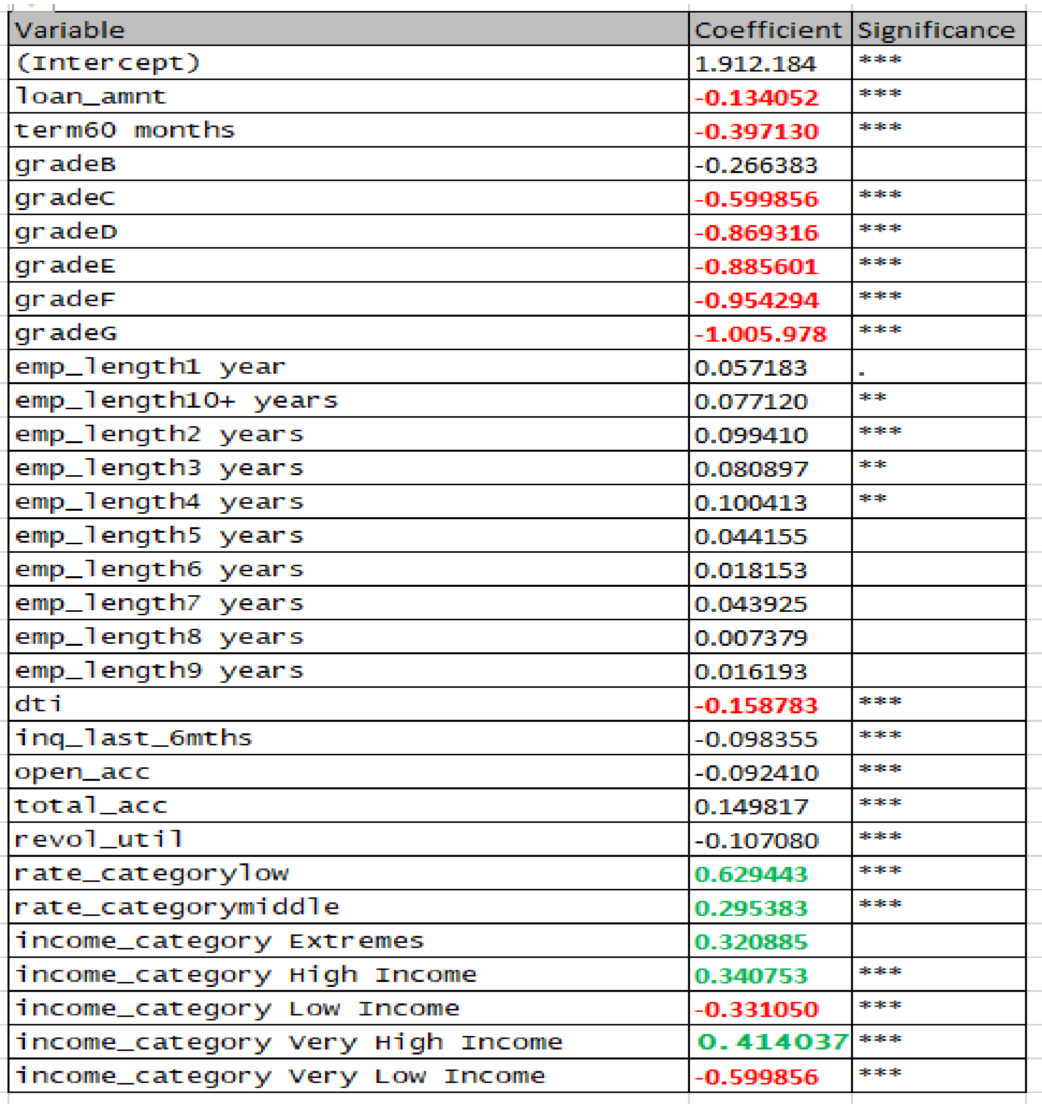


Figure 44: ROC Curves comparison after resampling

However, despite the fact that K-NN and Decision Trees are best fit models compared to Logistic Regression for the prediction of defaulters, there is a major disadvantage of both algorithms if we consider the results in a business perspective. Logistic Regression – given that it produces an equation between the target variable and the explanatory predictors- has the ability to give insights on the analyst in terms of which variables and how they contribute in the prediction of the final status of a loan. In other words, Logistic Regression explains in a much more comprehensive way the results and it is very easy to be interpreted and give insights to the top management in terms of how every explanatory variable influences the final outcome.

The coefficient of each explanatory variable "explains" the analyst how the corresponding variable influences the final probability estimation. A variable with positive coefficient and significant statistical power indicates positive influence on the calculation of credit score; in more simple words, a positive coefficient means incremental probability or greater credit score fir the predefined acceptance threshold. On the other hand, a negative coefficient means influences the final prediction in a reductive way. This is a perspective that – in many cases- is of greater importance for the decision maker than the accuracy and sensitivity of a model solely.

The coefficients of the explanatory variables in the Logistic Regression Model implemented in our analysis are presented in **(Table 17):**



Variable	Coefficient	Significance
(Intercept)	1.912.184	***
loan_amnt	-0.134052	***
term60 months	-0.397130	***
gradeB	-0.266383	
gradeC	-0.599856	***
gradeD	-0.869316	***
gradeE	-0.885601	***
gradeF	-0.954294	***
gradeG	-1.005.978	***
emp_length1 year	0.057183	.
emp_length10+ years	0.077120	**
emp_length2 years	0.099410	***
emp_length3 years	0.080897	**
emp_length4 years	0.100413	**
emp_length5 years	0.044155	
emp_length6 years	0.018153	
emp_length7 years	0.043925	
emp_length8 years	0.007379	
emp_length9 years	0.016193	
dti	-0.158783	***
inq_last_6mths	-0.098355	***
open_acc	-0.092410	***
total_acc	0.149817	***
revol_util	-0.107080	***
rate_categorylow	0.629443	***
rate_categorymiddle	0.295383	***
income_category Extremes	0.320885	
income_category High Income	0.340753	***
income_category Low Income	-0.331050	***
income_category Very High Income	0.414037	***
income_category Very Low Income	-0.599856	***

Table 21: Coefficients of Explanatory variables in Logistic Regression Model

Clear as it may be, some explanatory variables influence the calculation of credit score in a positive way. Examples of those variables are:

- Rate category → value "Low"
- Rate category →value "Middle"
- Income Category → value "Extreme"
- Income Category → value "High"
- Income Category → value "Very High"

In simple words, an analyst concludes that an applicant with High income who applies for a loan with Low interest rate, has greater probability to be consistent to their obligation towards Lending Club.

On the other hand, there are also negative indicators towards the acceptance of an application:

- Grade (B or worse)
- dti
- Income Category → value "Low"
- Income Category → value "Very Low"

As discussed in the theoretical framework of this Master Thesis, the strategy that is finally decided by the top management, is a combination of parameters and assessments and thus, a model that is better "communicated" to the top management maybe be much more preferable in the long run by a model that over performs the other by a few percentage points of accuracy.

The last part we will present in this section is the conclusions made towards the research questions that influenced the author of this Master Thesis. We tried to give answers in three questions related to the theoretical framework introduced in this Thesis:

1. What is the process of constructing a classification model for credit risk scoring?
2. What incremental value does a classification model for predicting the probability of default add to a financial institution?
3. Which classification algorithm suits better for the purpose of credit scoring calculation and risk assessment?

We will answer in every question in detail:

1. ***What is the process of constructing a classification model for credit risk scoring?***

We have answered in this question in a very explicit way. We tried to introduce a framework according to the existing bibliography through which we described an end to end classification modelling process. The modelling process involves many important steps before the implementation of the final algorithms. Initially, the need of a model development should be yield from the financial institute and specific strategy should be made available to the analyst in terms of the final goals of the model implementation. In other words, first should be decided the reasons why a classification model should be developed; is it for minimizing the risk exposure towards possible defaults? Or the top management searches for more profit opportunities in the existing financial sector. Maybe it would be a combination of both perspectives.

Next, the credit analyst should gather all appropriate data and start a process of data preparation. That involves, variable selection, descriptive statistics, deriving new KPI's not available in the initial data and finally start the modelling process. Those steps are very important for the final predictive power and results of the model as described in the existing bibliography introduced in this Master Thesis (Anderson, 2007), (Abdou & Pointon, 2011) (Brown & Moles, 2016) (Bhatia, Sharma, & Burman, 2017) (Hand & Henley, 1997) (Louzada, Anderson, & Guilherme, 2016). The next step of data preparation and data preprocessing is model implementation. Several techniques are introduced, both statistical and machine learning oriented (Altman, 1968), (Bhatia, Sharma, & Burman, 2017), (Abdou, Pointon, & El-Masry, 2008), (Hand & Kelly, 2002), (Moody & Haydon, 2018), (Wang, Wang, & Lai, 2005), (Unknown Author, 2018).

3 techniques have been put under analysis in this Master Thesis; Logistic Regression, Decision Trees and K Nearest Neighbors. We have chosen those 3 algorithms between all algorithms introduced for two reasons; first the author of this Thesis decided to implement the most common used algorithms derived by the bibliography and second because Neural Networks and Support Vector Machines are algorithms that need much greater computing power that was available at the author the time this Master Thesis was written.

Finally, the results of each model are evaluated through 3 common metrics: Confusion Matrix – a table that includes both actual and predicted values in a tabular way- AUC and ROC, which are metrics that compare the ability of a model to be able to distinguish between a positive class observation and a negative class observation.

2. ***What incremental value does a classification model for predicting the probability of default add to a financial institution?***

We have discussed in the theoretical framework of this Master Thesis, the importance of credit evaluation in a financial institutions' credit management decisions (Abdou & Pointon, 2011). This process includes collecting, analyzing and classifying different credit elements and variables to assess the credit decisions. The quality of bank loans is the key determinant of competition, survival and profitability. So, the main target of banks' decision making relies on the early identification of the quality of loans in terms of the probability to default. One of the most important kits, to classify a bank's customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring.

One of the main aims of the credit risk manager is to analyze different dimensions and aspects of an applicant's profile in order to assess whether or not this individual would be on time at their obligations towards the financial institute or not. In other words, discriminate applicants in two big, mutual exclusive categories; those lenders that will pay their loans in time, and those that default on their loan within given time. However, due to the fact that no manager is able to "predict" a future outcome, he is not aware of "the type" of a client beforehand and needs to decide whether to give a loan based on a set of variables provided by the client themselves (application data), third party data providers (credit agencies' data) or historical behavior of the customer (data on previously taken loans) (Herasymovych, 2018).

A set of decision models and their underlying techniques that aid lenders in the granting of consumer credit (Gup & Kolari, 2005). Thomas (Thomas L. C., 2000), indicates that those techniques tent to "decide" who will get credit, how much credit and most important, what operational strategies will enhance the probability and the amount of profit of the borrowers to the lenders.

A classification model should be regarded as a "safety net" for a financial institute. If we consider how difficult would be to differentiate between tens, even hundreds of available customers' details those that eventually should be considered as true performance indicators of the financial ability of an individual, the implementation of a well-trained classification algorithm which will aid decision making in the first line of recognizing potential risks is of high importance.

Without a proper model implementation in the initial application steps, the decision makers would only rely on guesses; that means, future status would be only assessed as the average number of defaulters through the available historical data a financial institute may have access to. Therefore, decision making would be derived from arbitrary metrics, with catastrophic results. Classification algorithms developed in this Master Thesis gave almost 70% of correct recognition of future defaulters. This number seems to be small if concerned towards the high level of risk exposure that the other 30% of incorrect predictions may result to, but it is much better than the default 50% of a random guess.

3. ***Which classification algorithm suits better for the purpose of credit scoring calculation and risk assessment?***

We have tested 3 different classification algorithms in terms of accuracy and sensitivity. As described in **8.4,** K- Nearest Neighbors over performed Logistic Regression and Decision Trees in terms of correctly predicting the percentage of actual defaulters. However, we argued towards the selection of Logistic Regression as the results of this algorithm are much more easily interpreted and communicated to the top management for the final decision making and the strategy setting of the financial institution. Further research should be made on the implementation of Support Vector Machines and Artificial Neural Networks in Lending Club datasets, as those algorithms propose better classification results according to the bibliography (Haardle & Schaafer, 2003), (Huang & Chen, 2007), (Landajo, de Andres, & Lorca, 2007). Additionally, different variables should be tested as explanatory indicators and new derived KPI’s may give even better insights or generate other statistically significant explanatory variables that may result in better predictive sensitivity.

Chapter 9
Conclusions

Three classification models were trained on the initial dataset which was highly imbalanced towards the minority class observations (defaulters) the percentage of whom was less than 20%. The first results demonstrated very low predictability power of true defaulters (25-30%). SMOTE resampling technique has been applied in the initial dataset with significant correction of sensitivity in all 3 models. True Negative proportion has almost tripled using SMOTE resampling due to the resistance of the method in overfitting compared to resampling or random sampling techniques.

K-NN over performed Logistic Regression and Decision Trees both in accuracy and sensitivity. Logistic Regression though is proposed as an easier interpreted model both for the analysts and the decision makers given that it reveals positive and negative influence between the target and the explanatory variables. No obvious differences are found between the models in terms of accuracy but in terms of sensitivity, the 3% difference between K Nearest Neighbors and the other two models ranks the machine learning algorithm on top of the list towards the prediction of true defaulters.

The maximum sensitivity of 72% our models succeeded, indicate that the problem of correctly recognizing a future defaulter is more difficult that it may seem to be and the risk exposure for a financial institution remains high. Additionally, Artificial Neural Network models and Support Vector Machine should be put under comparison with the 3 models developed in this Master Thesis as researchers often argue in favor of ANN and SVM models, in terms of succeeding greater accuracy and sensitivity (Shawe-Taylor & Cristianini, 2000), (Hui & Sun, 2006), (Abdou, Pointon, & El-Masry, 2008), (Backman & Zhao, 2017).

Bibliography

1. Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, 1275-1292.

2. Abdou, H., & Pointon, J. (2011).
Credit scoring, statistical techniques and evaluation criteria: A review of the literature . *University of Salford*.

3. Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 589-609.

4. Anderson, R. (2007). The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. *Oxford University Press*.

5. Backman, D., & Zhao, J. (2017, July). *Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling.* Retrieved from www.moodysanalytics.com: https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling

6. Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 1582-1594.

7. BBC. (2016, January 16). *AI pioneer Marvin Minsky dies aged 88.* Retrieved from bbc.com: https://www.bbc.com/news/technology-35409119

8. BBC. (n.d.). *BBC.* Retrieved from www.bbc.com

9. Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71-111.

10. Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 3302-3308.

11. Bennett, D. A. (2001). How can I deal with missing data in my study? *AUSTRALIAN AND NEW ZEALAND JOURNAL OF PUBLIC HEALTH*.

12. Bhatia, S., Sharma, P., & Burman, R. (2017). Credit Scoring using Machine Learning Techniques. *International Journal of Computer Applications*, 161.

13. Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, pp. 1145-1159.

14. Breiman, L. (1996). Bagging predictors. *Machine learning*, 123-140.

15. Brown, K., & Moles, P. (2016). *Credti Risk Management.* Edinburgh: Heriot-Watt University.

16. Chawla, N. V. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pp. 321-357.

17. Chung, H., & Tam, K. (1992). A comparative analysis of inductive learning. *Intelligent Systems in Accounting, Finance and Management*, 3-18.

18. Crook, J. N., Edelman, D., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 1447-1465.

19. Cukier, K., & Schoenberger, V. M. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*.

20. Damrongsakmethee, T., & Neagoe, V.-E. (2017). Data Mining and Machine Learning for Financial Analysis. *Indian Journal of Science and Technology*, 10.

21. De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference proceedings. 1644*, pp. 97-104. AIP.

22. Durant, D. (1941). *Risk elements in consumer installment financing.* New York: National Bureau of Economic Research.

23. Ertekin, S., Huang, J., & Bottou, L. (2007). Learning on the border: active learning in imbalanced data classification. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* (pp. 127-136). ACM.

24. Fawcett, T. (2005). An introduction to ROC analysis. *Science Direct*, pp. 861-874.

25. Fisher, R. (1986). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 179-188.

26. Fletcher, D., & Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. *Information and Management*, 159-167.

27. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning 29*, 131-163.

28. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 137-144.

29. Gately, E. (1996). *Neural Networks for Financial Forecasting: Top Techniques for Designing.* New York: John Wiley & Sons, Inc.

30. George, G., Haas, M. R., & Pentland, A. (2014). Big Data and Management: From the Editors. *Academy of Management Journal*, 321-326.

31. Gup, B., & Kolari, J. E. (2005). *Commercial banking: The management of risk.* John Wiley & Sons Incorporated.

32. Haardle, W., & Schaafer, D. (2003). Predicting corporate bankruptcy with support vector machines. *Humboldt University and the German Institute for Economic*.

33. Hadi, J. H., & Shnain, A. H. (2015). BIG DATA AND FIVE V'S CHARACTERISTICS. *International Journal of Advances in Electronics and Computer Science*, 16-23.

34. Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit scoring: a Review. *Journal of the Royal Statistical Society: Series A*, 523-541.

35. Hand, D. J., Sohn, S. Y., & Kim, Y. (2005). Optimal bipartite scorecards. *Expert Systems with*, 684-690.

36. Hand, D., & Kelly, M. (2002). Superscorecards. *IMA Journal Management Mathematics*, 273-281.

37. Hand, J. D., & Jacka, S. D. (1998). *Statistics in Finance.* London: Arnold Applications of Statistics.

38. Herasymovych, M. (2018). Master Thesis: Optimizing Acceptance Threshold in Credit Scoring using. University of Tartu: Faculty of Social Scienses, School of Economics and Business Administration.

39. Hodges, A. (2016). *Alan Turing: Creator of modern computing.* Retrieved from bbc.com: https://www.bbc.com/timelines/z8bgr82

40. Hogenboom, M. (2016, January 20). *The most beautiful equation is Byes Theorem.* Retrieved from www.bbc.com: http://www.bbc.co.uk/earth/story/20160120-the-most-beautiful-equation-is-bayes-theorem

41. Huang, C.-L., & Chen, M.-C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications,* pp. 847-856.

42. Hui, X. F., & Sun, J. (2006). An application of support vector machine to companies' financial distress prediction. *In International Conference on Modeling Decisions for Artificial Intelligence* (pp. 274-282). Berlin, Heidelberg: Springer.

43. Ipsilantis. (2018). Lectures for module Advanced Quantitative Methods for Enterprise Risk Manger Masters Degree.

44. Johnson, R. W. (1992). Legal, social and economic issues in implementing scoring in the US. *Credit Scoring and Credit Control*, 32.

45. Kenton, W. (2018, May 31). *Credit Risk*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/creditrisk.asp

46. Kewat, P., Sharma, R., Singh, U., & Itare, R. (2017). Support vector machines through financial time series forecasting. *Electronics, Communication and Aerospace Technology , International conference of IEEE*, (pp. 471-477).

47. Kohavi, R., & Provost, F. (1998). Confusion matrix. *Machine learning*, 271-274.

48. Landajo, M., de Andres, J., & Lorca, P. (2007). Robust neural modeling for the cross-sectional analysis of accounting information. *European Journal of Operational Research*, 1232-1525.

49. Lee, T., & Chen, I. (2005). A Two-Stage Hybrid Credit Scoring Model Using Artificial Neural. *Expert Systems with*, 743-752.

50. Lim, M. K., & Sohn, S. Y. (2007). Cluster-Based Dynamic Scoring Model. *Expert Systems with*, 427-431.

51. Louzada, F., Anderson, A., & Guilherme, F. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 117-134.

52. Mackenzie, J. (2018, December 19). *How Machine Learning and Big Data is Changing the Future of Credit Risk Management.* Retrieved from medium.com: https://medium.com/iveyfintechclub/how-machine-learning-and-big-data-is-changing-the-future-of-credit-risk-management-9a08aa398ec5

53. Madden, S. (2012). From databases to Big Data. *IEEE Internet*, 16.

54. Manyika, J., & Chui, M. (2011, May). *Big data: The next frontier for innovation, competition, and productivity.* Retrieved from McKinsey Digital: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

55. Marr, B. (2015, February 25). *A brief history of big data everyone should read.* Retrieved from World Economic Forum: https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/

56. Messier, J., & William, F. (1988). Inducing rules for expert system development: an example using default and bankruptcy data. *Management Science*, 1403-1415.

57. Mester, L. (1997). What’s the point of credit scoring? *Business review*, 3-16.

58. Michalski, R. S., Carbonell, J., & Mitchell, T. M. (1983). An overview of Machine Learning. In R. S. Michalski, J. Carbonell, & T. M. Mitchell, *Machine Learning: An Artificial Intelligence Approach.* California: TIOGA Publishing Cp.

59. Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, 603-614.

60. Mircea, G., & Pirtea , M. (2011). Discriminant analysis in a credit scoring model. *Recent advances in applied & biomedical informatics and computational engineering in systems applications*, 56-69.

61. Moody, H., & Haydon, D. (2018, August 20). *A Perspective On Machine Learning In Credit Risk.* Retrieved from spglobal.com: https://www.spglobal.com/marketintelligence/en/news-insights/research/a-perspective-on-machine-learning-in-credit-risk

62. Mukid, M. A., & Widiharih, T. (2018). Credit scoring analysis using weighted k nearest neighbor. *Journal of Physics*.

63. Murphy, K. (2018). Machine Learning: A Probabilistic Perspective . *Adaptive Computation and Machine Learning series*.

64. Musumeci, F., & Rottondi, C. (2018). An overview on application of machine learning techniques in optical networks. *IEEE Communications Surveys & Tutorials* .

65. Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems. *Journal of American Statistics Association*, 799-806.

66. Nikam, S. S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. *Oriental Journal of Computer Science and Technology*, 13-19.

67. NilSson, N. J. (1998). *Introduction to Machine Learning.* Stanford: Stanford University.

68. Norman, J. (2017). *The Williams Tube and the "Manchester Baby".* Retrieved from historyofinformation.com: http://www.historyofinformation.com/detail.php?entryid=858

69. Odom, M., & Sharda, R. (1990). A neural networks model for bankruptcy prediction. *Proceedings of the IEEE International Conference on*, (pp. 163-168).

70. Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of money, Credit and Banking*, 435-445.

71. Pampel, F. C. (2000). *Logistic regression: A primer.* Sage.

72. Power, D. J. (2014). Using ‘Big Data’for analytics and decision support. *Journal of Decision Systems*, 222-228.

73. Provist, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 51-59.

74. Reuters. (2014, December 4). AvantCredit Raises $225 Million From Tiger Global, Peter Thiel.

75. Royal Society. (2017). *Machine learning: the power and promise of computers that learn by example.* The Royal Society.

76. Safavian, S., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*.

77. Salzberg, S. (1997). “On Comparing Classifiers : Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, pp. 317-328.

78. Sandberg, M. (2017). Credit Risk Evaluation using Machine Learning. Linko¨ping University.

79. Sarlija, N., Bensic, M., & Bohacek, Z. (2004). Multinomial Model in Consumer Credit Scoring. *10th International Conference on Operational Research.* Trogir: Croatia.

80. Seen, K.-S., & Lee, T. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 127-135.

81. Shaw, M., & Gentry, J. (1990). Inductive learning for risk classification. *IEEE Expert*, 47-53.

82. Shawe-Taylor, J., & Cristianini, N. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge: Cambridge University Press.

83. Shin, K.-S., Taik, S., & Kim, H.-j. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications* , 127-135.

84. Sohn, S., Dong, H., & Yoon, J. (2016). Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing*, 150-158.

85. Soley-Bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis.*

86. Straseer, F. (2017, May 12). *When computers started ruling chess.* Retrieved from bbc.com: https://www.bbc.com/news/av/world-us-canada-39888639/how-a-computer-beat-the-best-chess-player-in-the-world

87. Sustersic, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications*, 4736-4744.

88. Swetz, J. A., & Robyn, M. D. (2000). Better decisions through science. *Scientific American*, pp. 82-87.

89. Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of. *International Journal of Forecasting*, 149-172.

90. Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications.* Siam.

91. Tian, X., & Deng, F. (2004). A credit scoring model using Support Vector Machine. *Fifth World Congress on Intelligent Control and Automation*.

92. Unknown Author. (2018). *Discriminant Analysis*. Retrieved from Statistics Solutions: https://www.statisticssolutions.com/discriminant-analysis/

93. Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley.

94. Wang, Y., Wang, S., & Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, 820-831.

95. Ward, J. S., & Barker, A. (2013). *Undefined By Data: A Survey of Big Data Definitions.* School of Computer Science, University of St Andrews, UK.

96. Yan, X. (2009). Linear Regression Analysis: Theory and Computing. *World Scientific*, 1-2.