Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Diploma Thesis

Loan Risk Predictive Modelling on Lending Club Data

Elena Tumanov

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Elena Tumanov

Informatics

Thesis title

Loan Risk Predictive Modeling on Lending Club Data

Objectives of thesis

Diploma thesis deals with the prediction of the loan risk and the chances that the loan will not be repaid . The final goal is to build a loan risk model based on data provided by Lending Club that would help investors to predict the chances that borrower will not repay the loan.

Methodology

The analysis will be based on Lending Club database. Several predictive methods will be used to predict if the loan will not be repaid and how investors could use these prediction models when deciding about their investments. The main concerns of investors is how to maximize their investments and minimize the risks. The risk at Lending Club is higher than in the banks since deposits are not protected. Loan risk models will be built for investors using supervised machine learning methods that would help investors to predict if a borrower will not repay the loan based on data provided by Lending Club. There are several important steps of data mining that I will follow in this project. These steps include: business understanding, data understanding, data preparation, model building and testing and evaluation.

The proposed extent of the thesis

60 – 80 pages

Keywords

Big data, data mining, predictive modelling, exploratory analysis, loan risk, lending club, peer-to-peer, lending, credit scoring, FICO.

Recommended information sources

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

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

FICO(2018). Learn About The FICO® Score and its Long History.[online] Available at: http://www.fico.com/25years/ [Accessed 18.01.2018]

Hadley Wickham, Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 1st Edition. Canada: O'Reilly, 2017. ISBN 978-1491910399.

Investopedia(2018). Peer-To-Peer Lending (P2P). [online] Available at: http://www.investopedia.com/terms/p/peer-to-peer-lending.asp [Accessed 18.01.2018]

Lending Club(2018). Lending Club Statistics. [online] Available at: Club:https://www.lendingclub.com/info/statistics.action [Accessed 18.01.2018]

Mandi Woodruff(2014). Here's what you need to know before taking out a peer-to-peer loan. [online] Yahoo Finance. Available at: http://finance.yahoo.com/news/what-is-peer-to-peer-lending-173019140.html [Accessed 20.01.2018]

NISBET, R. – MINER, G. – ELDER, J. Handbook of statistical analysis and data mining applications. Amsterdam: Amsterdam, 2009. ISBN 978-0-12-374765-5.

Roger D. Peng(2016). Exploratory Data Analysis with R. [online] Available at: https://bookdown.org/rdpeng/exdata/ [Accessed 10.02.2018]

Valerie Rind (2016).Pros and Cons of Peer-To-Peer Lending. [online] GOBankingRates. Available at: https://www.gobankingrates.com/personal-finance/5-perks-peer-to-peer-lending/ [Accessed 20.01.2018]

Expected date of thesis defence

2019/20 WS – FEM (February 2020)

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 15. 1. 2019

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 5. 2. 2019

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 27. 11. 2019

Declaration

I declare that I have worked on my diploma thesis titled "Loan Risk Predictive Modelling on Lending Club Data" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the diploma thesis, I declare that the thesis does not break copyrights of any their person.

In Prague on 28.11.2019

Elena Tumanov

Acknowledgement

I would first like to thank my supervisor Ing. Tomáš Hlavsa, Ph.D. for all of his help through this process. Additional thanks to Lending Club for making their data public and in this way facilitating such endeavours. I am especially indebted to my friends and family for their warmth and support.

Predictive Modelling on Lending Club Data

Abstract

In the lending industry, the loans are provided by investors to borrowers in exchange for the repayment with interest. In case the borrower repays the loan, the lender would have a profit from the interest. However, in case the borrower fails to repay the loan, the lender loses money.

Predicting the risk, if the borrower can be unable to repay the loan is the most common problem faced by the lenders. There is an immense risk while making a decision on a loan approval. Therefore, the goal of this diploma thesis is to gather data about loans from the Lending Club which is a peer-to-peer lending platform based in USA, and use machine learning techniques in order to extract meaningful information and predict if a borrower would be able or not to repay the loan, if the borrower would be Charged-off or not.

In this diploma thesis several Machine Learning models such as Random Forest, Logistic Regression and K Nearest Neighbors are trained to determine if the borrower has the ability to repay its loan. Furthermore, the performance of the models was also analysed. As a result, Logistic Regression model is chosen as the most optimal predictive model and it is likely that the Interest Rate and the Term significantly influence the forecast.

Keywords: Big data, Predictive Modelling, Exploratory Analysis, Loan Risk, Peer-to-Peer Lending, Linear Regression, K-Nearest Neighbors, Random Forest.

Prediktivní modelovaní ve Lending Club databáze

Abstrakt

V odvětví půjček, investoři financují dlužníky výměnou za splácení úroků. V případě, že dlužník splácí půjčku, půjčovatel má z úroku zisk. V případě, že dlužník nezaplatí úvěr, půjčovatel ztratí peníze.

Předvídání rizika, pokud dlužník nemůže být schopen splácet půjčku, je nejčastějším problémem, kterému půjčovatel čelí. Při rozhodování o schválení úvěru existuje obrovské riziko. Cílem této diplomové práce je proto shromáždit údaje o půjčkách od půjčovacího klubu, který je platformou vzájemného poskytování půjček se sídlem v USA, a pomocí technik strojového učení extrahovat smysluplné informace a předvídat, zda by dlužník byl schopna nebo nesplacená půjčka, pokud by byl dlužník splacen nebo ne.

V této diplomové práci je vyškoleno několik modelů strojového učení, jako je Random Forest, Lineární Regrese a K Nearest Neighbors, aby dlužník měl schopnost splácet svou půjčku. Dále byl analyzován výkon modelů. Výsledkem, model Logistické Regrese je vybrán jako nejoptimálnější prediktivní model a je pravděpodobné, že úroková sazba a termín významně ovlivňují prognózu.

Klíčová slova: Big data, Prediktivní modelování, Průzkumná analýza, Úvěrové riziko, půjčky typu Peer-to-Peer, Lineární Regrese, K-Nearest Neighbors, Random Forest.

Table of content

List of pictures

List of tables

List of abbreviations

P2P – Peer-to-Peer

LDA - Latent Dirichlet Allocation

kNN - K-Nearest Neighbors

SVM - Support Vector Machine

PCA - Principal Component Analysis

SGD - Stochastic Gradient Descent

ROC - Receiver Operating Characteristic

AUROC - Area Under the Receiver Operating Characteristic

GLM – Generalised linear model

SVR – Support Vector Regression

GPR – Gaussian Processes Regression

TP – True Positive

TN – True Negative

FP – False Positive

FN - False Negative

IQR – Interquartile Range

KS – Kolmogorov - Smirnov

1 Introduction

The guarantee offered by Business Intelligence and Big Data is well known in the past years. Companies claim that making data-driven decisions could revolutionise the business world but are not fully successful at it yet. It's known that descriptive analytics is useful and needed to understand data but in order to get real value for data analysis, companies must have the future in mind. The real value of the Big Data can be obtained from the continuous execution of the predictive analytics.

Business Intelligence has been used for data analysis for many years and it was offering valuable results but that's only a fraction of value. In the past Big Data Analysis was mostly focused on the Exploratory Analytics which is the most simplistic way of analytics a lending company can use. Exploratory Analytics has on its base a large data set, the big data, and is taking into consideration the past events. Instead of making predictions on how the things can be changed and improved based on the learnings from the data, Exploratory Analytics only summarises the data. It's excellent for showing the progress what is made against the goals the companies set and consider ways to improve their strategy, so it serves more like a reporting tool.

The future of Analytics nowadays is Predictive Analytics which takes into account the patterns found in the data in order to identify opportunities and risks. Advanced statistical modelling techniques, data mining and machine learning are ways of predicting the future events from the past and present events. It allows financial institutions to use statistical models on loan portfolios in order to decide the final outcome.

Predictive Analytics like Data Mining and Machine Learning techniques are broadly used for the loan default prediction, the peer-to-peer lending loan type which is becoming more and more popular. These platforms allow borrowers to borrow money from different investors which are private and exclude the traditional lending from banks.

Lending Club is built based on a peer-to-peer lending model which is remodelling the banking system. It's located in San Francisco, California with the main purpose to enable borrowers to make loan requests and investors to find these requests based on their preference. The range of the loan amount is from $1,000 to $40,000 for a period of 36 either 60 months. The data is available on the website and it's grouped by the issue year. On Lending Club borrowers describe the purpose of their loan and provide required information about their financial situation. The advantage for the lenders is the income which is being

generated in the interest form and is often greater than the amount of interest earned using traditional ways.

The main goal of this diploma thesis is to identify a predictive model that will present the highest accuracy at predicting if the loan will be Fully Paid or Charged Off which may be helpful for potential investors while evaluating the loan applications and deciding on their investment. The second goal which will be covered is to find out and analyse the relationships and associations between the attributes of the applications in order to determine other valuable knowledge for the potential investors. The main concerns of the potential investors are how to maximise the investments and minimise the risks as the risk at Lending Club is higher than in the banks because deposits are not protected in any form.

2 Objectives and Methodology

2.1 Objectives

The main objective of the current thesis is to perform an extensive analysis in order to identify the best models which would provide potential investors knowledge about the borrower, if the loan is going to repaid or not.

The Diploma Thesis can be divided in 2 main parts:

1. Perform Exploratory Analysis which is used to find the relationships and correlations between the attributes which may be helpful in the decision-making process. Individual attributes analysis, bivariate attribute analysis and comparative analysis will help to identify valuable knowledge about the data which is affecting the final outcome.
2. Build multiple predictive models and identify which one performs the best and would provide potential investors knowledge about the borrower, if it's going to repay the loan or not. By building 3 models a comparative analysis will be done in order to evaluate parameters and performance and identify the best one which has potential and deserves to be improved.

2.2 Methodology

The objectives of this Diploma Thesis can be achieved only by a in-depth review of the Big Data and its characteristics. In Chapter 3 a proper research on the lending topic is done in order to get a necessary business understanding and a financial background which is required in order to get a better understanding of the business model. The business model is described in the same chapter among with the lending process. The literature analysis and practical use cases from this chapter help to take smarter decisions in the practical part. The last part from the literature review contains the data preprocessing process which and the description of the predictive models which will be used to achieve the objectives, such as Random Forest, K-Nearest Neighbors and Logistic Regression.

The Chapter 4 deals with data understanding at first, Data Dictionary analysis, exploration and the quality assessment. That is the Lending Club dataset which is available on Kaggle.com (George, 2019). The data used for this analysis starts from the beginning of their activity in 2007 to 2018Q4 and it has 2260701 records and 151 features. The overal quality is good and considering that the dataset is big, there is a chance remove the entries which have more than >50% missing data. After, the data preparation proces follows. In this

part the selection of the required data is done. Some variables contain data gathered after the loan has been issued. These are removed and also observations with >50% missing data are removed from the dataset and decisions are made regarding what loans to keep. The remaining variables which will be left will be analysed in detail, transformed and formatted where needed. These preparations are made in order to make the variables usable in the model building stage.

The Chapter 5 includes the modelling. As the selection of the appropriate modeling technique was done in the Chapter 3, this part starts with grouping the loans to "Fully Paid" and "Charged Off", and then hypothesis tests are used to compare the distributions of each feature. Pearson Correlation is going to be used in order to detect the linearly correlated features and if these variables can be excluded from the modelling as this can lead to misleading results. After, the Feature Engineering part comes where as a first step the data is split into Train and Test samples with a ratio of 7:3. After the missing values are imputed with a help of an imputer model which is trained on the training data and used to fill in the missing values in both the test and training samples. The missing values are filled in with the median data from training sample in order to avoid the data leakage problem. Further, the feature scaling is performed as some of the models take into account the distance measures between observations and scaling allows them to learn. Finally, the Machine Learning algorithms are implemented followed by evaluation and comparison. The modelling techniques which are studied and used are Logistic Regression, Random Forest and K-Nearest Neighbors Algorithm. Evaluation is using a ROC curve on the training sample, cross-validated AUROC score analysis on test sample, confusion matrix and the performance measures computed from the confusion matrix on both the training and test sample in order to describe the performance of the implemented models, compare the models and decide on the best model.

3 Literature Review

3.1 The Big Data Concept

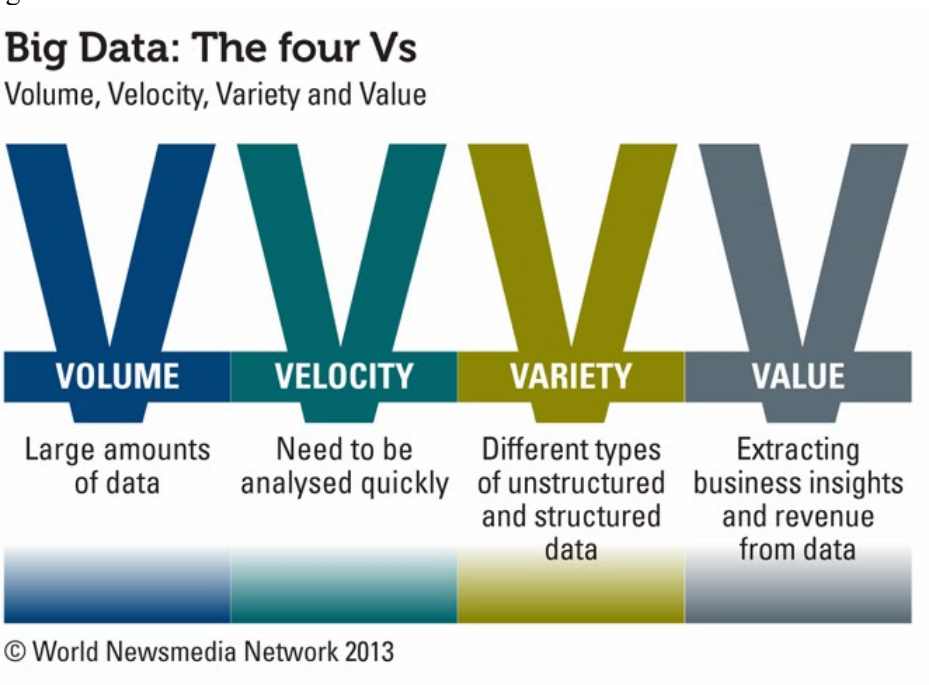
Big Data is a generic name for structured and unstructured data of huge volumes that is efficiently processed using scalable software tools. Such tools appeared on the market in the late 2000s and became an alternative to traditional databases and business intelligence solutions. (Semenov, 2019)

Big data analysis is performed in order to obtain new, previously unknown information. Such discoveries are called insights, which means assumptions and general understanding.

When talking about Big Data, it's worth mentioning the VVVV rule – four V criteria that big data should meet, shown in Figure 1:

1. Volume - data is measured by the size of the physical volume of documents.
2. Velocity - data is regularly updated, which requires continuous processing.
3. Variety - a variety of data can have heterogeneous formats, to be unstructured or partially structured.
4. Veracity – means reliability and data trustworthiness.

Figure 1 Big Data: The four Vs



Source: World Newsmedia Network, 2013

3.1.1 **What is Big Data?**

Recently, the term Big Data has become increasingly popular. Obviously, the emergence of this concept is somehow connected with the sharp increase in the amount of information available for analysis. Indeed, in most articles on the topic of Big Data, the story of a new technology begins with a discussion of the problem of the exponential growth of data. Social networks, mobile devices, indicators of various kinds of equipment, all kinds of business information, scientific research - these are just a few sources of the formation of "big data".

Is the mankind able to analyse so powerful streams of heterogeneous information? International experience suggests that organizations that have begun to use Big Data technology can gain a significant competitive advantage, as they can make decisions based on timely, relevant, accurate and complete information. Opportunities associated with the "big data" analysis can drastically affect the redistribution of companies in the market: organizations that do not seek to use the benefits of using "big data" technology will have to push aside. The fact is that the volume of all data increases exponentially: in 2011, the amount of generated information reached 1.8 zettabytes, in 2012 - 2.8 zettabytes. By 2020, this figure will multiply many times, shifting the mark of 40 zettabytes.

By "Big Data" one should understand not only conditional petabytes of accumulated information, but also a set of technologies aimed for storage, processing and services. It would seem that working with such a volume of information is unlikely outside of science, where the Large Hadron Collider generates 1 petabyte of data per second. But this is only the visible part of the question. If we dig a little deeper, it becomes clear: the economy and Big Data are interrelated structures. According to the Table 1 it's clear that for the analysis of big data, statistical, clustering and regression methods are used, as well as machine learning and artificial neural networks. Its' necessary to mention that the use of big data to predict or explain patterns is restricted because the result is highly dependent on the models that are used for analysis. (Hurwitz, et al., 2013)

Big Data is no longer a fashion trend, but a technology, without which modern business cannot compete in the market.

Table 1 Big Data, Data Mining and Machine Learning

Function	**Task**
Big Data - the actual raw data arrays.	Storing and managing large amounts of constantly updated information.
Data mining - the process of processing and structuring data, the stage of analytics to pattern identification.	Structuring a variety of information, search for hidden and non-obvious links to bring to a common denominator.
Machine learning - the process of machine learning based on the detecting relationships in the analysis process.	Analytics and forecasting based on the processed and structured information.

Source: (Semenov, 2019)

3.1.2 **Big Data Analysis in Financial Sector and P2P**

The term Big Data was first introduced in 1977 and since then has been the most mysterious tool in the economy. Some people believe in the extraordinary efficiency of this technology, but others are sceptical. Is this a powerful decision or waste of time when it comes to finances? Big data in economics is a significant amount of unstructured data. In contrast to statistics, the collection of information is not following any specific parameters and it does not have a predetermined goal. All the available indicators are considered for analysis, to determine the existence of relationships and to be used to improve the efficiency of the company.

Must be taken into account that the cheapness of collecting and storing Big Data is offset by the need to involve professional analysts. Fortunately, technology is gradually crowding out human resources, making it possible to operate with large amounts of information relatively inexpensively. The growth of data volume allows to fully take into account all the factors and risks affecting the development of a particular company. A simple example is the activity of a credit institution. Using just statistics, the bank can understand whether the borrower will repay the loan at the moment, but it is impossible to predict anything. Big amount of information will allow to build a predictive model and understand how high is the risk that the borrower will not pay the loan. Moreover, enough information can provide an opportunity to act ahead of the curve. After a proper analysis of more important parameters, the loan can be offered to a “reliable” borrower, who is ahead of other borrowers. The only question is the correct interpretation of how quickly and correctly a big

data scientist can turn informational chaos into an “informational diamond”. According to specialists, if all the developing countries of the world would use the Big Data with the same intensity as the advanced countries like Singapore, the Netherlands, the USA, Germany, the global GDP will grow by 13%. (Manyika, et al., 2014)

Why Big Data is important for the economy? What is the actual use of Big Data? While individual companies are “bragging” about the millions they have saved, let's find out what the introduction of this tool gives to the economy as a whole.

These are some advantages: (Akopian, 2016)

- Reduction of financial risks - up to 60% in the banking sector;
- Increased flexibility - eliminating loss of income due to insufficient management awareness;
- The desire for greater automation - reduced labour costs, expensive human resources are gradually replaced by less expensive software;
- Increased customer satisfaction - when creating any product, from diapers to train stock, many more factors are taken into account;
- Thanks to the introduction of technology, the economy is developing more actively, all sectors are becoming more competitive. Accordingly, the level of well-being of people employed in this field is increasing.

In P2P lending Big Data is being used to analyse trends and help to understand who the borrowers are, their behaviour and the trends in order to minimise the risk of investment. Unfortunately, is not true that in order to be a successful investor, the investor should be able to understand and manipulate big data. Big Data is a helpful resource, but it cannot do everything.

There are 4 important factors about Big Data in P2P Lending: (Lustman, 2013)

1. Big Data is considering the Past Performance only

The historical information which is available on the loan listing shows how the borrower has behaved in the past. The fact that the history shows that he is paying his bills on time doesn’t mean that he will pay the loan on time, but it increases the probability. Also, if he has a stable job, this adds value. Historical information is reflective in the past but is not predictive itself.

2. The Loan Pool is not big

Taking Lending Club as a practical example, from the beginning in 2006 to November 2012 they issued the first $1 billion in loans. After one year the total of $3 billion of loans

was issued. As of September 2018, $40 billion of loans were issued. The number of issued loans is increasing, and this means that 2/3 of the total loans are recent, 1 year or less. Some sources write that there is no meaningful data but still the most meaningful data from this historical record is the likelihood of an early payment default.

3. Big Data Isolates dependent attributes

Data Set is highly available and allows to research very little details like Debt Consolidation loans for 600 FICO score for the Monday loans. One of the main benefits of the big data is that it allows to study every variable which is good and bad in the same time. The problem is that it isolates these variables and these variables are dependent on each other and isolation devalues it. Credit score for example is dependent on many factors and these things are factored while filtering by credit score is done, so the isolation is affecting the quality of the filter selection and is redundant.

4. Correlation isn't necessary Causation.

If two variables are related this doesn't mean that the change in the first one will affect the second one. For example, the CA has the highest rate of default and also the highest number of issued loans. In case it's removed the pool of available loans will be reduced significantly and means that all the good loans that meet some filtering standards will be also excluded. The impact on the data can be anywhere depending on multiple factors.

Big data analysis is only a useful tool which places investors who do this ahead of other investors in the P2P lending but still, the awareness of the risks that are out there is very crucial.

3.2 P2P Lending

3.2.1 Background

P2P lending is usually referred to as "peer-to-peer lending", "peer lending" or "person-to-person lending", "social crediting ". This method is connecting borrowers and investors directly, without using traditional financial institutions like banks and credit unions as an intermediary. Commonly, P2P lending is carried out through particular online platforms where the user can act both as an investor and as a borrower. Most of the loans issued under such services are unsecured private loans. But in some cases, it is possible for the legal entities to take part in this process. P2P lending is different from the traditional lending and presents a high-risk, because most of the time it is impossible to conduct high-quality scoring and to check the borrower's credit history. Correspondingly, the rates for

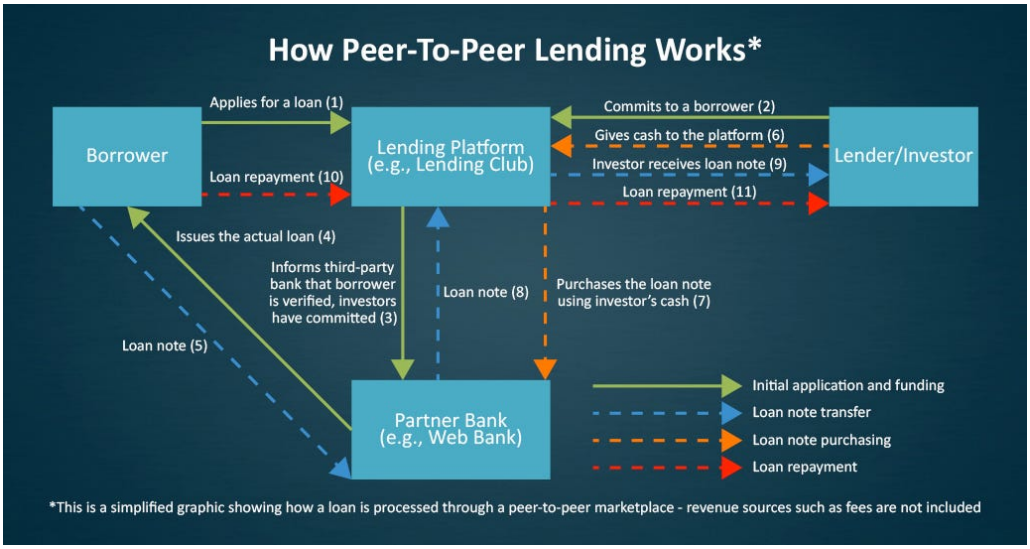
such loans are rather high. The risk of non-repayment can be reduced by issuing a multiple small loan to different borrowers.

P2P credit rates can be either fixed or determined based on a reverse auction. In the latter case, the potential borrower sets the maximum level of the rate at which he agrees to take out a loan, and potential investors bargain among themselves, offering money at a lower rate. Intermediary companies which are running P2P lending, receive income from fixed payments by borrowers or as a small percentage of the amount of loans taken from both the borrower and the investor.

Zopa (Zone of Possible Agreement) - the first company which introduced the P2P lending online - was created in the United Kingdom in 2005. (Wikipedia, 2019) The first P2P credit companies appeared in USA in 2006 - Prosper and Lending Club. Other P2P credit platform include Funding Circle, Trustbuddy and Thincats.

3.2.2 **How P2P Lending Works**

Figure 2 How P2P Lending Works



Source: Business Insider

The Peer-to-peer Lending is shown in the Figure 2 and it works as following. The borrowers apply for different reasons for loans. The main scope of the loan applications on Lending Club and Prosper is credit consolidation. The application is done by providing some relevant personal information through filling in a form on the platform. The information consists of loan amount, loan purpose, term, occupation, monthly income, credit rating and the debt-to-income ratio. The next step is to get and agree on the interest rate. Platform

assesses the entered information and decides on the interest rate for the loan. Borrower has to approve the interest rate in order to have the loan listed on the platform so the investors can see and take it. After this step is done the loan becomes available to the investors and they can browse the loan information and take a decision of investment, to invest or not and how much.

The most popular subject on P2P is how to increase the possibility of the Fully Paid loans and what are the most important factors. It's known that unverified variables play a less significant role than the verified ones when it comes to determine if it's worth to invest a loan or not. Borrowers who disclose more information about themselves choose to pay less interest rate. The overall image affects the decision of the lenders to fund the loan. Male borrowers are more likely to get loans funded than female borrowers. Many social factors have an important role in the lending process. Investors make decisions on lending and the exact amount based on the provided information by the borrower. The ultimate goal of lending money through P2P platforms is to mitigate the default risk and get high return back. Investors prefer high return loans which have also a high default risk.

Platforms are charging the borrowers a loan origination fee once the loan is successfully funded and investors a service fee for managing the instalment payments which come from the borrowers. Borrowers are entitled to pay monthly instalments till the moment when the loan reaches the maturity, they also can pay ahead all key payments with a service fee. Investors can sell holding loans with a certain discount on the trading system provided by the platform which provides more flexibility to investors. Some loans use to default in early stages of instalment payments and cause a big lose. (Kagan, 2019)

3.2.3 **P2P Lending Advantages and Disadvantages**

P2P lending platforms are gaining popularity. The benefits of investing with the prospect of a high income are quite a few: (Lustman, 2017)

- The most decisive advantage that distinguishes "social" lending sites from all other options of financial investment is high profitability. Due to the fact that practically anyone who is denied by a legal organization can receive money in debt, lenders have the right to raise interest rates for lending money, which makes P2P lending a profitable investment tool.

- On P2P platforms there are no such high limits for the minimum amount of investment. Often, the minimum amount starts from $25, which means that even small savings can generate income.
- P2P platforms conduct their activities online, so it eliminates the need to visit different institutions. The income can be earned without leaving the home, using a computer or a smartphone.

P2P lending has its own drawbacks, which can eliminate all the advantages: (Lustman, 2017)

- No guarantee of return. A platform that provides the opportunity to invest the money cannot distinguish the unreliable borrower. Due to the fact that the funds used in lending are the personal funds of borrowers, P2P platform does not bear any risks, and therefore is not too interested in the return of money. Sometimes there is an additional option of registration of insurance against default. Often, the cost of insurance can reach high amount, but it can protect against even greater spending in the future.
- Tax deductions. There are sites that independently control the payment of tax deductions from the receipt of income and make payments for the client. But most often, the control of taxes of individuals remains on their conscience. Evasion of payments threatens with a fine and even punishment.
- No early return on investment. When investing, it's clear that the money may not be able to return at any time. At best, funds can be received back at the end of the term of the contract, and at worst - with delays or even through the courts.
- The risk of financial fraud. High profitability and confusing working conditions of the organization are direct signs of a financial pyramid that can be eliminated at any time. To return the money in this case is very difficult, and maybe even impossible.

Verifying a P2P platform can protect the investor from financial loss.

Advantages and disadvantages for the borrower: (Rind, 2016)

- For borrowers who wish to get a small amount of debt, P2P platforms can become a lifesaver.
- The amount available for debt on P2P platforms can significantly exceed the amount of loans online or consumer loans at banks.
- Many companies that specialize in P2P loans prefer to work online, so getting the necessary amount of credit on the card will not be a problem. Also, as with the

issuance of a loan on the card, the response on the application does not need to wait long.

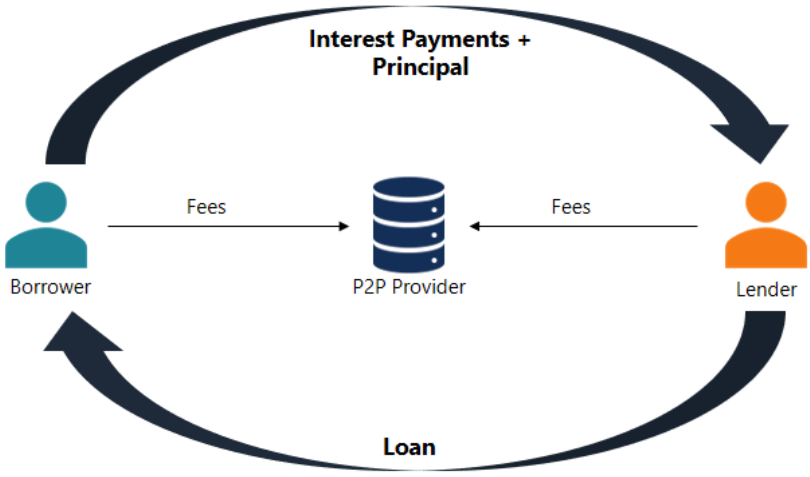
- Due to the fact that P2P lending is still not controlled by the state, the data on the receipt of money in debt will remain unnoticed and closed.
- On the lending platforms, it's easy to get money without a permanent job, as well as take a loan being unemployed. Certificate of employment doesn't need to be attached to the profile.

P2P has a negative side: in case of a conflict situation between the lender and the borrower, no one can protect the interests of the ordinary citizen. Problems will have to be solved through the court, and the proof of payment or receipt of money will be requested from the platform.

3.2.4 **P2P Business Models**

In this section the business models used by P2P lending platforms in USA will be introduced as this is helpful to understand the Lending Process which may vary from one to another. In the USA the P2P business models are quite similar which are shown in Figure 3. Lenders browse on the platforms through the posted loans and decide on investment. In this case the platform is just an intermediary part and is determining the interest rate, document checking, interest payment and is responsible in some way for risk taking.

Figure 3 P2P Business Model



Source: Corporate Finance Institute

Two main actors, Prosper and Lending Club have some differences: (Cunningham, 2019)

1. Loan type - Prosper is offering only personal purpose loans, $2000 - $35000 while Lending Club also business loans with a maximum of $300000 and personal from $1000 to $35000. The maturities are different from one to another, but both provide 3 and 5 years loans, Lending Club has also the 1 year option.
2. Interest Rate - the interest rate is determined by the platform according to the information which is filled in by the borrower and determines the credit quality. Both platforms have a fixed cap and floor value for the interest rate for the different rates and grades. The interest rate in the same credit category is different because of the different credit rating logic.
3. Credit Scoring - the credit score is provided by both platforms as a indicator of a loan risk. They have 7 categories, Prosper start from HR which is the worst to AA which is the best and Lending Club starts from G which is worst to A which is best.
4. Origination Fee - both platforms are intermediary between the borrower and investor and they earn money from the fees paid by the borrowers. The minimum and maximum are the same but there can be a difference depending on risk categories.
5. Affiliate & Referral Programs - the Prosper has an affiliate program which has the scope to attract more lenders and borrowers from referrers and to pay for it. Lending Club has also such a program, but details are not available.
6. Notes Trading Platform - provided by both platforms which allows investors to trade their notes.
7. Early Repayment - in order to avoid paying the monthly interest rate, without any penalty the borrowers can pay the remaining repayment.
8. Interest Auction - P2P lending platforms usually regulate the interest rate for loans according to the information from the borrowers. Prosper has introduced the interest rate auction in which investors can compete for the best loans, but it stopped in 2011.

3.2.5 **Credit Scoring**

There are various Credit Scoring definitions. Credit scoring is a method of classifying borrowers into different groups to determine their creditworthiness and risk level. It's a mathematical or statistical model by which a bank or in our case lending platform determines the likelihood of a timely repayment of a loan by a specific potential borrower. Credit scoring is used in consumer express lending for small amounts and is based on analysis of data on

loan repayment by certain groups of borrowers obtained from studying a large number of credit histories. The use of credit scoring allows to simplify the work of the loan officer who reviews applications for credit and increases the speed of their consideration.

A great definition of Credit Scoring has. (Anderson, 2007) He states that the term Credit Scoring is made up of two parts which are credit and scoring. Credit from Latin Credo means trust and the Scoring is about using analytical methods in order to rank something for an easier differentiation. Nowadays Credit means get now and pay for it later and Scoring in this context is assigning a grade what describes the case. The combination of Credit and Scoring expresses the idea that Credit Scoring means using statistical models for the transformation of the data to a numerical score which describes the probability of the borrower's default.

Credit Scoring concept became popular because of the technological growth. It caused the occurrence of the more advanced credit scoring methods. Credit Scoring belongs to the risk management category in financial institutions being an indispensable fact of it, which helps to get maximum profit from the clients. This is possible only when the borrower's default rate is minimum which can be achieved by decreasing the informational imbalance between lenders and borrowers. Borrowers always can estimate their ability to pay the credit, information that is not available to lenders. Lenders know only the information provided by the borrowers in the application process about the loan and themselves. A credit scoring model is predicting the creditworthiness based on the loan application, but this can be obtained only if the credit scoring model is trained on similar loan applications which are finished, and the results are known.

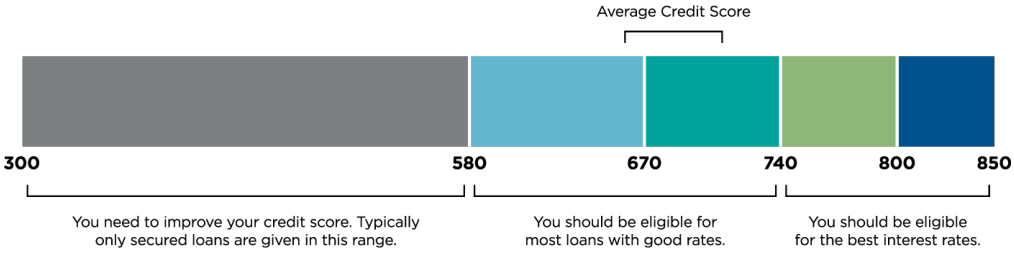
The purpose of using credit scoring systems are is to increase in the loan portfolio by reducing the number of unjustified refusals of loan applications, to reduce the level of loan defaults, to improve the accuracy and speed up the procedure for assessing the borrower, to collect data on borrowers, to reduce the amount of reserves for possible losses on credit obligations, etc.

The credit scoring procedure consists in assigning certain points to the borrower in accordance with the questionnaire filled out by him, developed by appraisers and bank credit inspectors, in order to determine the borrower's credit risk level. In terms of the amount of points scored automatically, the credit scoring system provides recommendations for deciding whether to approve or deny issuing a loan. In most credit programs, the scoring can

automatically decide on 85-90% of loan applications. In most cases, the decision issued by the credit scoring system on the credit application will be decisive.

One of the most popular credit scoring methods is the one introduced by the Fair Isaac Corporation, FICO-score, Figure 4. The FICO-score method is popular in USA, invented in 1989 and being used from 1991 by three credit reporting agencies. (Hayes, 2019) According to FICO a potential borrower is rated on a scale from 300 to 850 points: above 690 - "excellent"; 650-690 – "medium standard level"; 640-650 - "very good"; 620-640 - "good"; 600-620 - "bad"; below 600 - "high risk".

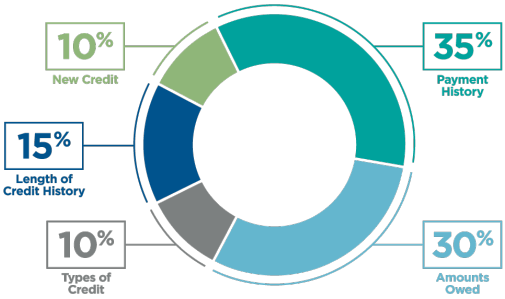
Figure 4 FICO credit scores range



Source: Community America, 2019

All credit reporting agencies use a different version according to the database structural differences which makes difficult to compare scores between the agencies. The exact calculation used by the reporting agencies is a secret, the only known thing are the type of factors influencing the score, shown in Figure 5.

Figure 5 Factors influencing the score



Source: Community America, 2019

According to the Figure 5, the Payment History has the most value in score calculation, with the weight of 35%. The Amounts Owed 30%, the Length of Credit History

15%, the Type of Credit and the New Credit weight 10% each. Consequently, the three credit reporting agencies mentioned above, because of the dominant position of the FICO-score and the inability to correlate the scores, decided to design in 2006 new credit rating score – Vantage Score. Payment history is whether or not the borrower pays the bills on time each month - affects the credit score more than anything else. The next most influential factor is the ability to maintain a mix of accounts like credit cards, auto, student loans, mortgage, etc. over time, followed by the ability to limit the amount of used credit, idealy keeping revolving balances under 30% of credit limits. After that, it's always a good idea to keep the total debt balance low. While these last two are considered less influential, opening too many new accounts too quickly, or taking more credit than actually needed, can still have an impact on the score.

Credit scoring in P2P is very similar to the one used in the regular financial institutions. Lending Club uses the FICO-score in order to do the initial scanning and estimate the interest rate. In the application process Lending Club asks for the relevant information and for the creditworthiness assessment. Some critical information like yearly income can be verified prior the loan approval or denial. In case the loan is approved, it gets assigned a grade from A to G. Each grade is divided in five subgrades, from 1 to 5. The subgrades have a corresponding interest rate which take into account macroeconomic factors. (Lending Club, 2019)

3.2.6 **Lending Process at Lending Club**

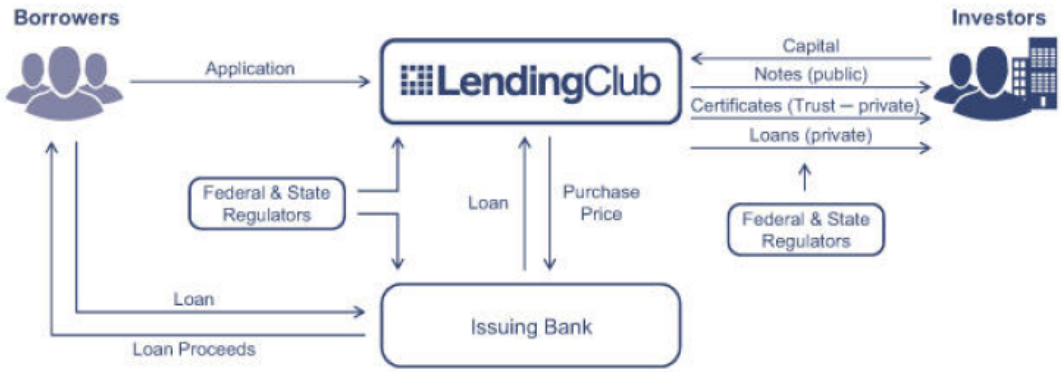
The awareness of a lending process helps to understand the principles of a successful loan application. Also, the data generation process should be taken into consideration. Some characteristics like FICO score has changed few times. The basic loan application process did not change in time, but the application requirements and options are different.

The main requirement which is also the minimum condition for the loan application is the FICO-score value. The minimum value is 600 but the higher is the FICO-score, the higher is the creditworthiness. If the minimum requirements are met, the borrower is asked to provide more personal and loan related information. (Woodruff, 2014)

First, a self-reported annual income is required to be provided followed be home situation with the options like mortgage, rent, own, other. The employment status and employment length are the next important factors but the second one is known based on the Social Security Number. Also, the borrower should provide the loan amount, loan purpose

and loan description. The loan purpose includes fourteen categories and is a required field, the loan amount is from $1000 to $35000 and the description is an optional field which is not filled in most of the applications. Based on the information provided by the borrower and the credit file information, the creditworthiness is determined using the Lending Club's credit scoring algorithm. The creditworthiness includes the credit grade and interest rate. After the scoring process ends, the loan offer is becoming available to the borrower. If the borrower agrees on it and approves it, the offer is becoming available for the investors on the Lending Club platform. Sometimes, in the course of the founding, Lending Club may ask the borrower to verify the self-reported annual income. In case that in the intervening time the loan is funded, the self-reported annual income information is not needed. Usually is known from the credit file if the verification is necessary or not. In case of the failure, the loan is removed from the Lending Club platform. (Loughlin, 2015) Process described in Figure 6.

Figure 6 Lending Process at Lending Club



Source: Harvard Business School

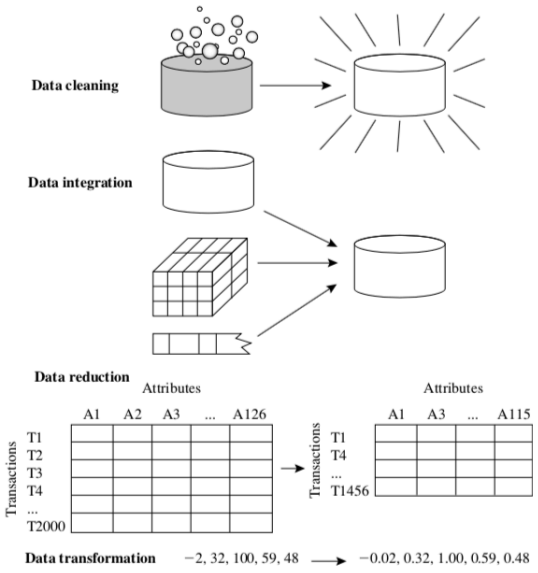
3.3 **Data Preprocessing**

When modelling real processes, "clean" data is a rare luxury. Real data usually contains noise and is unevenly distributed. Very often, the data is simply collected and sent to the model input, hoping that everything will work out. However, a preliminary data processing can save a lot of time and not only. Data preparation is a very important process which can take up to 80% of the time. Data quality is analysed based on the use requirements. Data completeness, consistency, accuracy, believability, timeliness, interpretability are the factors which compromise data quality. Most common data proprieties are incompleteness

and inaccuracies. Data may be inaccurate as a consequence of a collection method, computer or human errors at the data entry, users sometimes submit wrong data, transformation process, inconsistency in naming conventions or coding, inconsistent fields formats. Incomplete data is caused by unavailable attributes, missing data because it wasn't considered to be important, not recorded data because of misunderstanding or technical problems. The data quality is connected with the scope of the use of the data. If that data is not updated, timeliness can affect the data quality in a negative way. Interpretability is how easy the data is understood, and believability is how much is trusted by users.

As shown in the Figure 7, the most important tasks in data preprocessing are data cleaning, data integration, data transformation and data reduction. Data cleaning consists of filling in or deleting entries with missing values, adjusting noisy data, removing outliers and adjusting inconsistencies. Dirty data can cause inaccurate outputs and create confusion. Data reduction is reducing the size of the dataset which can still produce the same result. This is done by reducing the dimensions or numerosity. Distance based algorithms require data to be normalised because otherwise the distance measurements will outweigh. Techniques like concept hierarchy generation and discretisation are used to replace the attributes data with the higher conceptual levels which allow multiple abstraction levels. These three are forms of the data transformation. (Nisbet, et al., 2009)

Figure 7 Data preprocessing methods



Source: Introducing Data Science, 2016

Preprocessing can improve the data quality and in this way improve the model result. Detecting problematic data and adjusting it can save time and money later on.

3.3.1 **Data cleansing**

One of the subprocesses of the data science process is data cleansing. The goal is to remove errors in the dataset in order to get consistent data. Some common errors which require data cleansing are shown in a table Table 2.

Table 2 Common errors in datasets

Error type	**Solution**
Data entry mistakes	Manual disregards
White space	String functions
Impossible/unreal values	Manual disregards
Missing data	Remove/substitute
Outliers	Check and remove or substitute
Alteration from codification	Match or use change manually
Various units	Recalculate
Aggregation	Extrapolation or aggregation

Source: **(Cielen, et al., 2016)**

Missing data should be handled during the data cleansing process. There are more solutions to be considered during this process which are be selected based on the dataset type presented in the Table 3. The missing value observation can be ignored which is not very effective unless there are only few observations. If many observations have missing data, some important variables can be deleted, and this could affect the analysis. Missing data can be filled in manually which is too time consuming for large data sets. A global constant can be used to replace the missing values like “Unknown” but this can cause the model to consider the notation as a common value. The mean for normal data distribution or median for skewed data distribution can be used to replace missing values. Also, the most probable value can be used which is determined using regression, decision trees etc.

Table 3 Methods to handle missing data

Method	**Advantage**	**Disadvantage**
Remove values	Easy	Loose the observation
Set to null	Easy	Null values are not supported by all the modelling techniques
Input static value like mean or 0	Easy and the observation is saved not deleted	False outputs from the model
Input value from theoretical or estimated distribution	Does not affect to model almost at all	Execution is more complex Data is based on assumptions
Model the value	Does not affect to model almost at all	Dependence can be raised artificially Execution is more complex Data is based on assumptions Model can be too confident

Source: (Cielen, et al., 2016)

3.3.2 **Data transformation**

Certain models require the data to be in a certain format. The rapport between an input variable and an output variable isn‘t always linear. It can happen that there are too many variables and the number should be reduced because they don’t add additional information to the model. Too many variables in the model makes the model difficult to control, and some techniques don’t perform well when overloaded with too many input variables. (Cielen, et al., 2016)

3.3.3 **Combining data**

In order to combine data from two different data sets, two methods can be used. The first is joining: enriching an observation from one table with information from another table. The second is appending or stacking: adding the observations of one table to those of another table. When combining data, there is an option to create a new physical table or a virtual table by creating a view. The advantage of a view is that it doesn’t consume more disk space.(Cielen, et al., 2016)

Joining tables allows to combine the information of one observation found in one table with the information from another table. The focus is on enriching a single observation. To join tables, variables that represent the same object in both tables are used. These common fields are known as keys. When these keys also uniquely define the records in the table, they are called primary keys.

3.4 Predictive Analytics

Predictive Analytics is a method to learn from past experience represented by data in order to be able to predict the future experience for better decisions. The best way to predict the future is to look at the past events. The analysis is done with lot of data, statistic algorithms and assumptions. (Bari, 2016)

Predictive Analytics main components:

1. Data – the data part is not only about the quantity, an important point is to use this data in order to improve future decisions quality based on data quality.
2. Statistics – are mathematical methods applied to the data to deduce predictions.
3. Assumptions – key point in predictive modelling, assumptions define the model.

Predictive Analytics is very important for any organisation because it helps to solve challenging problems and discover new opportunities.

These are the main uses: (Abbott, 2014)

Reducing risk – credit scores are a well-known example of predictive analytics. It is used to assess a likelihood of default for borrowing or purchasing something. This is a number that includes relevant data about someone's creditworthiness and its generated by a predictive model.

Marketing optimisation – predictive models help organisations to interact with their profitable customers by attracting, retaining and growing. Analytics show the customer behaviour.

Operational improvements – organisations use predictive models to forecast and manage resources. Hotels predict guest number in different periods to increase the revenue, airlines set prices for the tickets. Predictive analytics is allowing to work more efficient.

Predictive analytics facilitates the usage of big data and enables the organisations to get benefits from the collected data.

Predictive Analytics applies on data statistics and machine learning to create predictive models which can forecast the future. The application of the statistical and machine learning techniques facilitates the creation of a quantitative prediction. (MathWorks, 2019)

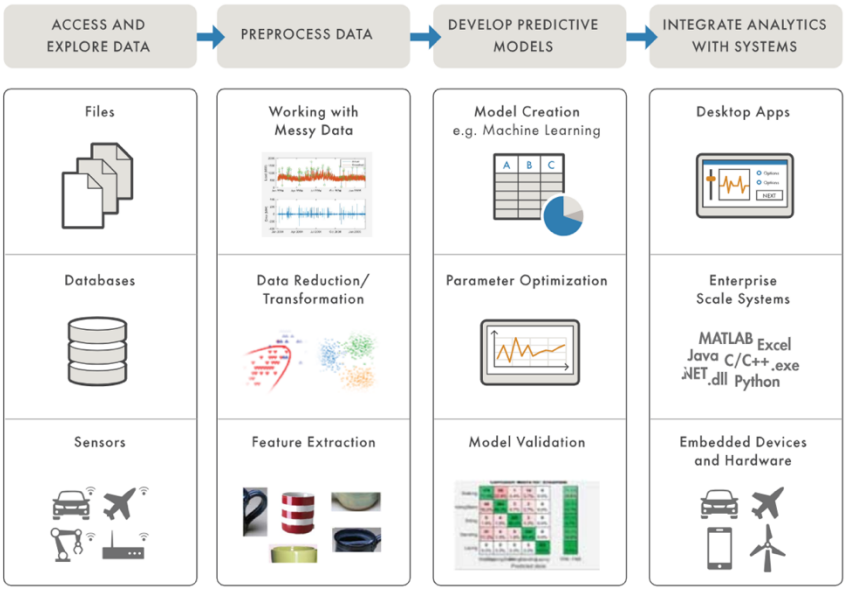
Predictive analytics starts with a goal and continues with the proces of combining data sets in models. The models generate outcomes which support the goal achievement.

Predictive Analytics application workflow has four steps shown in Figure 8:

1. Data collection from any available source and in any available format.

2. Data cleaning by combining data sources, merging files, eliminating entries with too much missing data, removing outliers, identifying data spikes.
3. Predictive model development using statistics or machine learning and based on the collection of the data. Iterations through the training data set are usually necessary to find the different and best approach. After the training completion it can be tested against the new data te see its performance.
4. Model integration into a forecasting system after the model forecasts accurately the load.

Figure 8 Predictive Analytics Workflow



Source: MathWorks, 2019

3.5 Machine Learning for Predictive Models

3.5.1 Machine Learning Overview

Machine learning is a predictive analytics branch and is the core principle behind predictive modelling. Predictive modelling is used to predict events or outcomes using computational and mathematical methods. The models forecast a future state which occur under the changes made to the model inputs. An iterative process is used to develop an accurate model. Training and test data sets are necessary for training, testing and validation. Two differences between machine learning and predictive analytics must be considered for

a better understanding of the future work. First, predictive analytics works on data only and needs to be refreshed if data changes while machine learning works in real-time and recalibrates models automatically. Secondly, predictive analytics needs humans to test the associations between the goal and the result, and machine learning doesn't.

"The field of Machine Learning seeks to answer the question "How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" (Mitchell, 2006)

"Machine learning is the science of getting computers to act without being explicitly programmed." (Stanford, 2019)

"Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world." (Copeland, 2016)

"Machine learning is based on algorithms that can learn from data without relying on rules-based programming." (Pyle & San José, 2015)

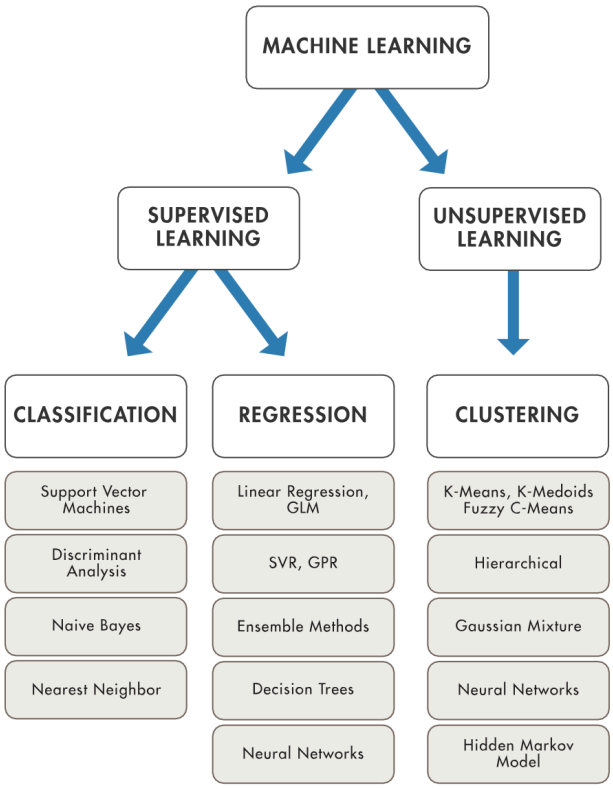
"Machine learning algorithms can figure out how to perform important tasks by generalizing from examples." (Domingos, 2012)

Multiple definitions and viewpoints on machine learning exist. In my understanding this is an intersection between theoretical computer science and practical data, and machines making outcomes out of data. The basic task is to build algorithms which use statistical analysis to predict an output. Machine learning is very simple, just an attempt to do something that even animals are capable of, to learn from experience.

3.5.2 **Machine Learning Algorithms**

Machine learning algorithms are divided into two categories, supervised and unsupervised. (Marsland & Stephen, 2011) According to the Figure 9, the supervised ones are can solve classification problems and regression problems. The classification problems can be solved with the Support Vector Machines, Discriminant Analysis, Naïve Bayes and Nearest Neighbor. The Regression ones with the Linear Regression, GLM, SVR, GPR, Ensemble Methods, Decision Trees and Neural Networks. The Unsupervised Learning handles clustering problems which are solved with K-Means, K-Medoids Fuzzy C-Means, Hierarchical, Gaussian Mixture, Neural Networks and Hidden Markov Model.

Figure 9 Machine Learning Algorithms



Source: MathWorks, 2019

Supervised learning needs human help to provide input, output and furnish the feedback about the accuracy during training. After the training completion, algorithm applies the learning on new data. Supervised learning uses regression and classification techniques for predictive models development. Regression techniques are used for continuous responses prediction. Usually used for data range or real number responses. Supervised learning is a good choice if a model needs to be trained to make predictions. (Marsland & Stephen, 2011)

Unsupervised learning is not trained with the outcomes data and uses the deep learning iterative approach for data reviews and conclusions. These algorithms are usually used for more complex problems. Clustering is the most popular technique user for exploratory data analysis to group data and identify patterns. Unsupervised learning is a good choice if data needs to be explored and a model is trained to find good data representations. (Marsland & Stephen, 2011)

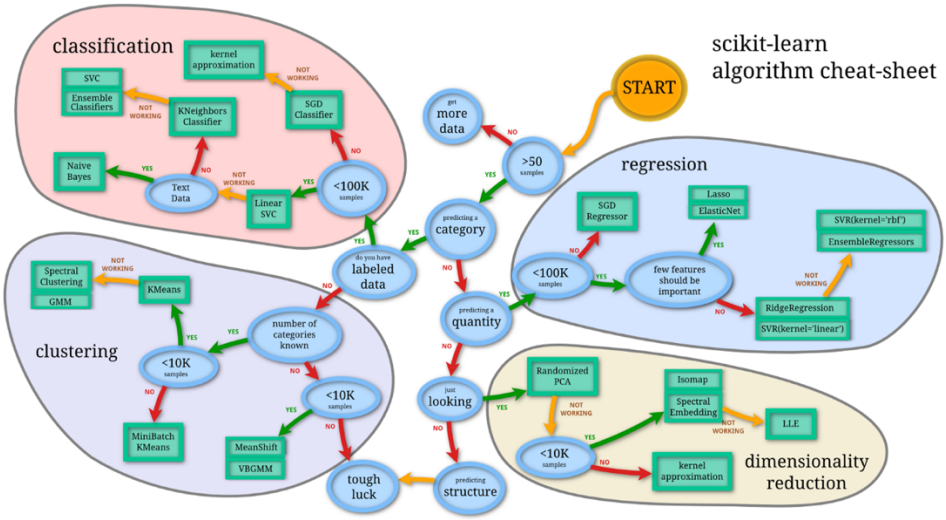
Further will be selected and examined the algorithms which are appropriate for the goal of this thesis.

3.5.3 **Model selection**

Three models are selected to be implemented in the practical part, Logistic Regression, Random Forest and K-Nearest Neighbors. The following factors were considered in the algorithm selection using as a guide also the Figure 10:

1. Size of the training set
2. Type of the problem
3. Training Time
4. Accuracy
5. Linearity
6. Number of parameters
7. Number of features

Figure 10 Scikit-learn algorithm cheat-sheet



Source: scikit-learn, 2019

3.5.4 **Logistic Regression**

Logistic Regression has the same aim as Linear Regression, but it predicts the qualitative response and it has as an output a categorical variable.

Linear Regression is first of all connected to statistics where its used as a model in order to understand the relationship between two numerical variables. Later it was implemented in machine learning. Is a linear model and it can be simple if there is one input variable or multiple when there is more than one input variable. Ordinary Least Squares is the most common algorithm used to test and train the model. The representation is simple, a

linear equation (1) which has input values. The solution is the output which is predicted for the input values. Both values must be numeric. Simple Linear Regression has linear relationship between the independent and dependent variable, with one independent variable.

In the following formulas the y is the response variable which denote $p = P(Y = 1)$, x_i are the predictors and β_i are the parameters of the model.

The equation which is used for modelling is:

$$y = \beta_0 + \beta_1 x_1 \quad (1)$$

Linear Regression (1) is used in order to find the best fit for 2 variables (β_0 and β_1).

Logistic Regression algorithm based on Linear Regression:

Linear Regression definition:

$$y = p(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad (2)$$

A model which would give a 0 and 1 output values is made based on a Sigmoid function and it‘s defined as following:

$$p = \frac{1}{1 + e^{-y}} \quad (3)$$

If y in (3) is changed with y in (2):

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (4)$$

If the dependent variable has two classes, then (4) is a conditional probability:

$$P(Y = 1 | x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (5)$$

If the numerator and the denominator of (5) are multiplied by $e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}$:

$$P(Y = 1 | x_1, x_2, \ldots, x_n) = \frac{e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (6)$$

On the other hand, there is:

$$P(Y = 0 | x_1, x_2, \ldots, x_n) = 1 - P(Y = 1 | x_1, x_2, \ldots, x_n)$$
$$= 1 - \frac{e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

After basic algebra operations:

$$P(Y = 0 | x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (7)$$

If (6) with (7) is divided, odds will be in favour that Y=1 $x_1, x_2, \dots, x_n$:

$$\frac{P(Y=1|x_1,x_2,\dots,x_n)}{P(Y=0|x_1,x_2,\dots,x_n)} = \frac{\frac{e^{(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}}{1+e^{(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}}}{\frac{1}{1+e^{(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}}} = e^{(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)} \quad (8)$$

Finally, with the logarithm of both sides in (8), we get:

$$ln(\frac{p(x)}{(1-p(x)}) = (\beta_0 + \beta_1x_1 + \cdots + \beta_nx_n) \quad (9)$$

As it was mentioned the quantity $\frac{p(x)}{1-p(x)}$ is called the odds in favor of p(x). Since this it is inside of the ln function, the domain of ln function is the interval (0, ∞). The odd quantity can have any value between 0 and ∞. If the value is very low, close to 0, it means that the p(x) probability is very low. If the value is very high, close to ∞, it means that the p(x) probability of is very high. The expression from the left side in (9) is named log-odds or logit. The right side is named linear predictor which is a combination of parameters $\beta_0 + \beta_1 + \ldots + \beta_n$ which have to be estimated and features $x_1, x_2, \dots, x_n$. The general method named maximum likelihood helps to obtain these unknown parameters. Maximum likelihood is a general approach used to fit many of the non-linear models. The logistic model can be fit in Jupyper Notebook in order to obtain the parameters $\beta_0 + \beta_1 + \ldots + \beta_n$ and get these coefficients. Other parameters are coefficients for features $x_1, x_2, \dots, x_n$ and β_0 parameter is intercept in the logistic model. The interpretation is very simple. If we consider that one of variables in the model x_k increases by 1, while everything else ceteris paribus, from (8) can be observed that the odds are multiplied by $e^{(\beta_k x_k)}$. In case β_k is negative, then $e^{(\beta_k)}$ is <1 which means that the odds will decrease as x_k increases. If β_k is positive, then $e^{(\beta_k)}$ is > 1 which means that the odds will increase as x_k increases. Once the unknown parameters are computed $\beta_0 + \beta_1 + \ldots + \beta_n$ it is very easy to make predictions applying the logistic regression model. If the probability of Charged off is predicted, in (4) all estimated parameters $\beta_0 + \beta_1 + \ldots + \beta_n$ together with values of the features obtained from data set should be plugged and this will give the probability of Charged-off. (Sertic, 2017)

Advantages and Disadvantages

It is a technique easy to implement and efficient to train. It's very used and is highly interpretable, does not require many computational resources, doesn't require the input features to be scaled and any tuning, outputs well-calibrated probabilities and is easy to regularize. A disadvantage is that it can't solve non-linear problems since its decision surface is linear. Also, is not from the most powerful algorithms and can be easily outperformed by more complex algorithms. It has a high reliance on the proper presentation of data which means that is not a useful tool unless all the important independent variables are already identified. Logistic Regression can predict a categorical outcome only because its outcome is discrete. It is also vulnerable to overfitting.

3.5.5 Random Forests

Random Forests corrects the overfitting problem. It uses multiple different decision trees which have corresponding different predictions. The result is a combination of the results from all the trees. The bagging ensemble algorithm is applied to the decision trees in order to reduce the variance and overfitting.

Having a training set and its outcomes, a random sample which replaces the training set is being selected by bagging a number of times. Decision trees are trained on that samples. The classification is made by the majority vote of the classification trees. In an algorithm that looks for one single hypothesis, usually three types of problems can occur: a statistical problem, the representation problem and the computational problem. These three problems can be partly overcome by ensemble methods, methods in which multiple models are used in combination with each other. (Dietterich, 1997) The Random Forest algorithm uses randomization instead of the successive reweighting. (Breimann, 2001) Random Forest combines the method of bagging with the random selection of features which have to be included. First the bagging technique is applied to tree learners which means that a random subset of the training data is taken and used to generate one tree per sample. After that a random set of attributes is selected at each split of each tree. This ensures that there is little correlation between the different member trees thus enabling high generalizability. Therefore, the Random Forest algorithm will be used in the context of this thesis.

Random forests algorithm:

1. Take a random sample of size n with a replacement.
2. Take a random sample of the predictor variables without a replacement.
3. Construct a regression tree using the predictors chosen in the random sample in step 2. Let it grow as much as it can. Do not prune the tree.
4. Pass the outside of the bag observations for this bootstrap sample through the current tree. Store the value or class assigned to each observation through this process.
5. Repeat steps 1 to 4 for a large number of times or the number of times specified (this is basically the number of trees one wants in the forest).
6. The final predicted value for an observation is the average of the predicted values for that observation over all the trees. In the case of a classifier, the final class will be decided by a majority of votes; that is, the class that gets predicted by the maximum number of trees gets to be the final prediction for that observation.

Advantages and disadvantages

Random Forests is an extremely flexible algorithm and can handle a variety of input data. Even without tuning of the hyperparameters it is able to produce quite good results. It allows to calculate the feature importance which lets us know what features are best able to separate target classes. Is also resistant to overfitting because the number of trees in the model is pretty high. A disadvantage is that, to run, it requires a lot of computational power and cost is higher the more trees are in the model.

3.5.6 KNN (k-Nearest Neighbors)

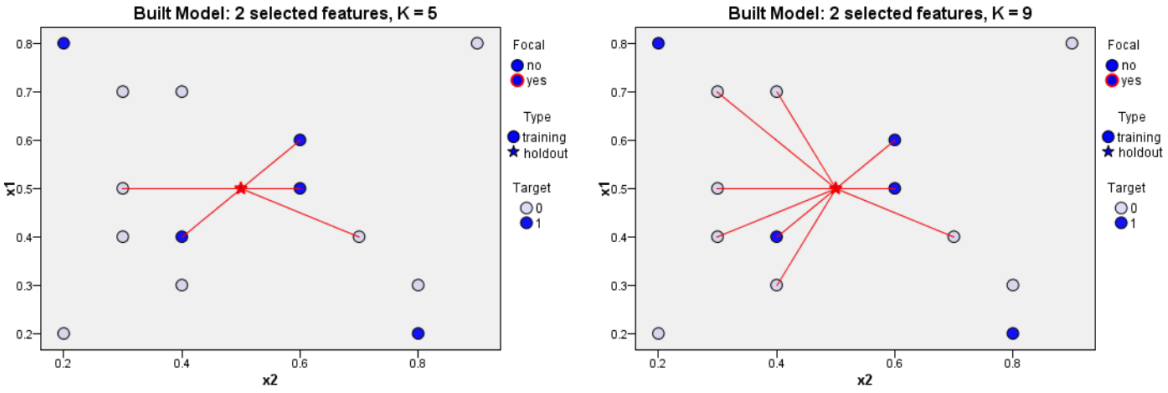
K-Nearest Neighbors is a supervised machine learning algorithm which can solve regression and classification problems. It assumes that similar things are close to each other and they exist in close proximity. The model includes the entire dataset, it doesn't require learning, which requires the training data to be consistent. Predictions are being made using the training dataset. A new instance is chosen for that. In order to find the K similar instances, neighbors, a search through the training dataset is done. For the K found instances the output variable is summarised. Using a distance measure, the identification of the most similar K instances from the training dataset is performed. The most common measure is the Euclidean

distance. The Euclidean distance is the square root of the sum of the squared differences between an existing point and a new one across all input values.

$$E(x, y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (10)$$

Other distance measures are the Manhattan Distance which is a distance between real vectors with the use of the sum of the absolute difference. The Hamming Distance which is a distance between the binary vectors. The Minkowski Distance which is a Euclidian and Manhattan Distance generalisation. The distance measure is selected based on the data proprieties. In order to get an accurate model experimenting with different metrics is recommended. In case the input variables have the similar type, the Euclidean distance must be used. In case they are not, the Manhattan distance is better. In order to identify the best K value, by algorithm tuning many different values should be used. If the dataset is big, the KNN computational complexity is also big. (Kuhn, 2018) In Figure 11, are shown two examples with K=5 and K=9.

Figure 11 K Nearest Neighbors Models Examples



Source: (Gandhi, 2018)

KNN Algorithm:

1. Load the selected dataset
2. Choose the number of neighbors and initialise K for each variable in the dataset.
3. Calculate the distance between the current example and the query example from the dataset.

4. Add the index of the example and the distance to an ordered collection
5. Sort the ordered collection indexes and distances in ascending order by the distances value
6. Select the first K entries from the sorted collection
7. Get labels of the selected entries
8. Regression - return the mean of the K labels
9. Classification - return the mode of the K labels

Advantages and disadvantages

An advantage is that it makes almost no assumptions on dataset because it is a non-parametric method. It is highly flexible, intuitive and simple to implement, and is providing good results in most cases. A disadvantage is that is sensitive to class-outliers. Observations that don't fit the general consensus can have a great impact on the predicted outcome. Another issue is that it does not do any feature selection and does not consider which features are more important than others. As a result of this it is very sensitive to irrelevant features. Computationally is a very expensive method. Being a memorization-based technique, it stores all the data in the training set. Also, it's expensive because it can take a long time to calculate all the distances needed to make the predictions. Is slower and slower the more data is included in the training set. The time spent by the model is in the testing phase, not in the training phase.

3.5.7 **Performance metrics**

The usual accuracy metric, based on misclassification, fails in the context of imbalanced data since a trivial classifier that just predicts the majority class performs well under this measure. For adequate measurement of the performance of imbalanced algorithms different metrics are necessary. This has been addressed extensively in the literature (He & Garcia, 2009). In Table 4 the relevant metrics for this thesis are defined.

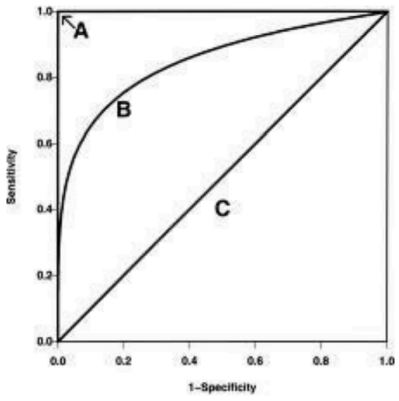
Table 4 The relevant metrics for assesing model performance

Measure	**Definition**
Recall	$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$
Precision	$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$
F-measure	$F - measure = \frac{(1+\beta)^2 * Recall * Precision}{\beta^2 * Recall + Precision}$
ROC graph	$Graph\ of\ sensitivity = \frac{TP}{TP+FN}\ versus\ 1 - specificity = 1 - \frac{TP}{TP+FP}$
Cost savings	Costs saved by implementing the model instead of random

Source: (He & Garcia, 2009)

The F-measure described in Table 4 combines both recall (completeness) and precision (correctness). The beta can take any non-negative value and is a way to emphasize one of the two. Normally a value of is selected. A limitation of the F-measure is that it's sensitive to data distributions (He & Garcia, 2009). The last one can also be applied to families of classifiers. The ROC-graph provides a graphical representation of the trade-offs between the benefits (the true positives), and the costs (the false positives). An example of such a curve can be seen in Figure 17. The line A represents a perfect predictor, the line C is a random predictor, and the line B a predictor with performance somewhere in between. The value of the area under the curve is also often used as a measure of accuracy and is called the 'area under curve'. This value varies from 0,5 totally random to 1 which is a perfect predictor.

Figure 12 ROC curve example



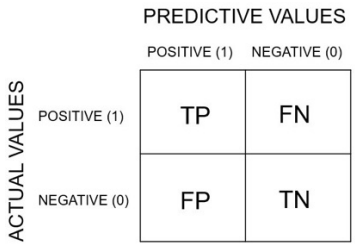
Source: (He & Garcia, 2009)

One of the most overall reliable measures for models is the ROC (Receiver Operating Characteristic) curve. The ROC-curve represents true positive rates against false positive

rates. The closer the curve is to left up corner, the model performance is better. In the real life situations when we compare different ROC-curves from different models, it is very frequent that is not clear which model performs better.

For classification models there are several measures of accuracy that which can be computed from the confusion matrix. These measures include: precision, accuracy, recall (true P rate), true N rate, FP rate and error. The confusion matrix is a contingency table of correct and incorrect classifications. The elements on the diagonal in the confusion matrix are called TP (true positives) and TN (true negatives), while elements on off-diagonal are called FN (false negative) and FP (false positive).

Figure 13 Confusion Matrix



Source: (He & Garcia, 2009)

The accuracy is described as percentage of correctly classified instances:

$$ACC = \frac{(TP+TN)}{(P+N)} \text{ (10)}$$

The error is described as percentage of incorrectly classified instances:

$$ERR = \frac{(FP+FN)}{(P+N)} = 1 - ACC \text{ (11)}$$

Precision or Positive Predictive Value (PPV) is described as fraction of real positive instances between all predicted positive instances:

$$PPV = \frac{TP}{(TP+FP)} \text{ (12)}$$

Sensitivity (recall, True Positive Rate, hit rate) is described as fraction of real positive instances (relevant instances) between all real relevant instances:

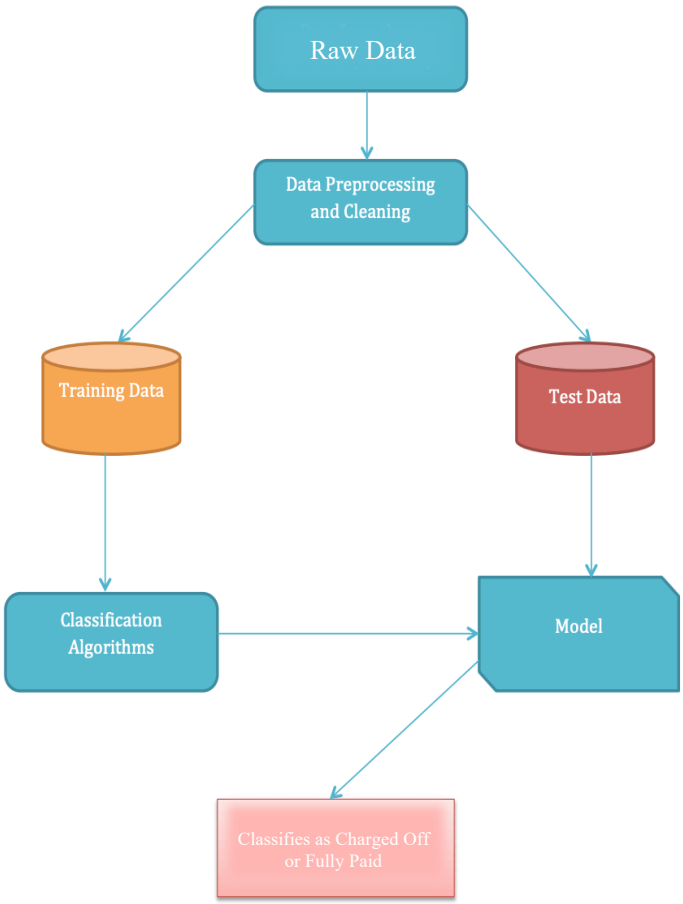
$$TPR = \frac{TP}{P} = \frac{TP}{(TP+FN)} \text{ (13)}$$

Specificity or True Negative Rate is described as a fraction of real negative instances among all real negative instances.

$$TNR = \frac{TN}{N} = \frac{TN}{(TN+FP)} \text{ (14)}$$

4 Practical Part

Figure 14 Workflow diagram



Source: Own

In Figure 18 the approach for the practical part of the thesis and all the following steps are illustrated.

4.1 Loan Data

The dataset which is used is a publicly available dataset on Kaggle (George, 2019) which contains data from Lending Club. The dataset has 2.5+ GB size, includes aprox. 2.2 million of loans from 2007 to 2018Q4, each of which has 151 associated features.

4.1.1 Variable Description

In the Appendix A are described the selected features which are investigated in the exploratory data analysis and used to build the predictive models.

4.1.2 **Initial Data Review**

The dataset consists of numerical and categorical features, and missing values. Because it has 151 features which is too much for the scope of this thesis, data engineering starts with dropping unnecessary features and then the remaining ones are investigated in more detail. The column loan_status is chosen to be predicted.

Value distributions for loan_status:

```
Fully Paid                                             1076751
Current                                                 878317
Charged Off                                             268559
Late (31-120 days)                                       21467
In Grace Period                                           8436
Late (16-30 days)                                         4349
Does not meet the credit policy. Status:Fully Paid        1988
Does not meet the credit policy. Status:Charged Off        761
Default                                                     40
NaN                                                         33
```

First of all are considered the features differences between the loans that have been completed, fully paid or charged off. The current loans, the loans which don't meet the credit policy, are defaulted, or don't have any status are not considered. Two types of loan status are considered and kept for the further analysis, with the assumption that they meet the credit policy. These are "Fully Paid" and "Charged Off" loans. A loan becomes "Charged Off" when there is no longer an expectation of further payments. "Charged Off" mostly occurs when a loan is 120 days or more late and there is no feasible expectation of a payment in order to prevent the it. Under certain circumstances, the loans can be charged off at an later or earlier date. (Lending Club, 2019) A loan is "Fully Paid" when has been fully repaid, either as a result of a prepayment or at the expiration of the 3- or 5-year year term. (Lending Club, 2019)

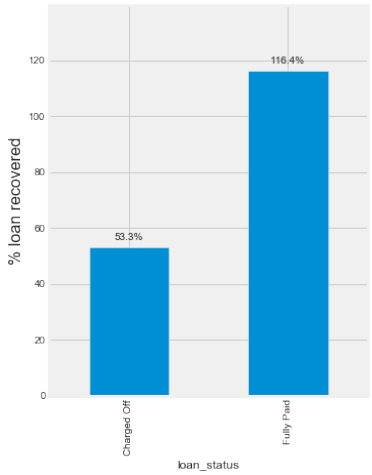
After dropping the following loan types: "Current Loans", "Does not meet the credit policy", "Defaulted" and "NaN", 1345310 loans are remaining.

The remaining loan statuses are:

```
Fully Paid     1076751                Fully Paid     0.800374
Charged Off     268559                Charged Off    0.199626
```

About 80% which is ~1m of the loans are under the status "Fully Paid," and 20% which is 199k are under the status "Charged Off". The data is unbalanced and should be handled in the cross validation. A business risk is considered any variable that increases percentage of default to more than 19.96%.

Figure 15 Recovered loans (%)



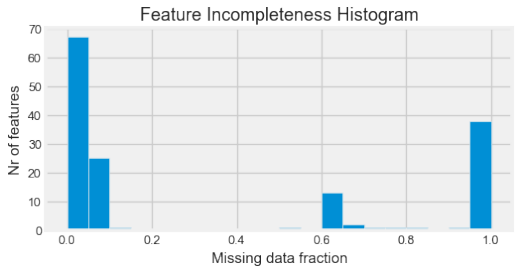
Source: own

The bar charts from Figure 15 show that Lending Club recovers only 53.3% of the loan amount in case the loans are "Charged Off" and on "Fully Paid" loans, they make 16,4% profit.

4.1.3 **Data Cleaning**

The dataset has 151 features, but not all of them are necessary in order to build the models. The definitions of the features are provided in the Lending Club "Data Dictionary" (Lending Club, 2019) and the definitions of the features considered in the exploratory data analysis can be found in Table 5. Features which have 50% and more missing data and data which is not available at the loan application point are excluded. Also, some additional adjustments like dropping the superfluous attributes, dropping highly correlated predictors and conversion of string to numerical values were done.

Figure 16 Feature Incompleteness Histogram



Source: own

Figure 16 shows that the data frame has a gap between features missing <15% of data and the ones missing >50%. Because it‘s very difficult to input data which has more than

50% missing values, 50% is the cut-off and all the features which have >50% missing values are dropped. No of dropped features is 58.

Lending Club Website and the Data Dictionary were examined in order to decide which features are available to investors. These are features from the loan application and the features added by Lending Club when the loan listing was accepted, such as loan grade and interest rate. As the prediction if a loan will be paid has to be made before making the decision to lend the loan, the following features are dropped:

```
['acc_now_delinq', 'acc_open_past_24mths', 'addr_state', 'annual_inc', 'applicati
on_type', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'chargeoff_within_12_mths',
'collection_recovery_fee', 'collections_12_mths_ex_med', 'debt_settlement_flag',
'delinq_2yrs', 'delinq_amnt', 'disbursement_method', 'dti', 'earliest_cr_line', '
emp_length', 'emp_title', 'fico_range_high', 'fico_range_low', 'funded_amnt', 'fu
nded_amnt_inv', 'grade', 'hardship_flag', 'home_ownership', 'id', 'initial_list_s
tatus', 'inq_last_6mths', 'installment', 'int_rate', 'issue_d', 'last_credit_pull
_d', 'last_fico_range_high', 'last_fico_range_low', 'last_pymnt_amnt', 'last_pymn
t_d', 'loan_amnt', 'loan_status', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', '
mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mt
hs_since_recent_inq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl
', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'nu
m_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_d
pd_24m', 'num_tl_op_past_12m', 'open_acc', 'out_prncp', 'out_prncp_inv', 'pct_tl_
nvr_dlq', 'percent_bc_gt_75', 'policy_code', 'pub_rec', 'pub_rec_bankruptcies', '
purpose', 'pymnt_plan', 'recoveries', 'revol_bal', 'revol_util', 'sub_grade', 'ta
x_liens', 'term', 'title', 'tot_coll_amt', 'tot_cur_bal', 'tot_hi_cred_lim', 'tot
al_acc', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit', 'to
tal_pymnt', 'total_pymnt_inv', 'total_rec_int', 'total_rec_late_fee', 'total_rec_
prncp', 'total_rev_hi_lim', 'url', 'verification_status', 'zip_code']
```

35 variables are remaining and each of them is analysed in order to decide if it should be dropped or not. The following variables are remaining:

```
['id', 'loan_amnt', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp
['id', 'loan_amnt', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp
_title', 'emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'is
sue_d', 'loan_status', 'url', 'purpose', 'title', 'zip_code', 'addr_state', 'dti'
, 'earliest_cr_line', 'fico_range_low', 'fico_range_high', 'open_acc', 'pub_rec',
'revol_bal', 'revol_util', 'total_acc', 'initial_list_status', 'policy_code', 'ap
plication_type', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mort_acc', 'pub_r
ec_bankruptcies']
```

In the exploratory part the following actions are done: visualization with count distributions, bar plot and Kernel Density Estimate plot; dropping useless features (e.g. "id"); cleaning and formatting (e.g. "earliest_cr_line", "emp_length"); transformation of the features (e.g. average of "fico_score", log transformation of "annual_inc"); selection of the most relevant features (e.g. "zip_code" or "addr_state", "sub_grade" or "grade").

4.2 Exploratory Analysis

Issue date (issue_d)

Data Dictionary: "The month which the loan was funded." (Lending Club, 2019) Because only the features which are available before the loan was funded are used to make predictions, issue_d will not be used in modelling but is kept for train/test sample split and it will be dropped after. Over the years Lending Club was giving more loans, from 2007 to 2015 the loans issued have risen significantly. Year of loan has no significant impact on likelihood of Charged Off.

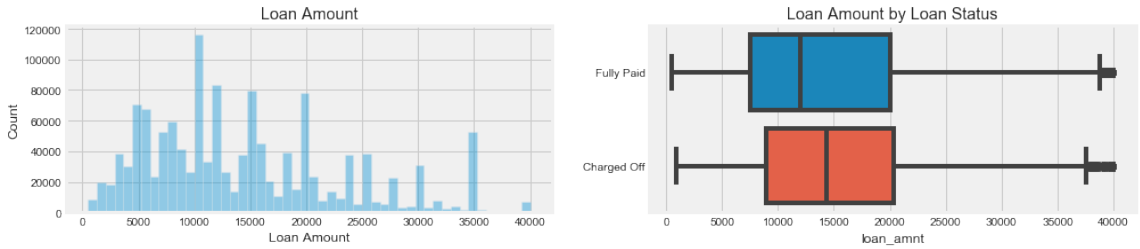
ID

Data Dictionary: "A unique assigned ID for the loan listing." (Lending Club, 2019) The "id" is a unique index which is not useful as a categorical variable for modelling and is dropped.

Loan Amount (loan_amnt)

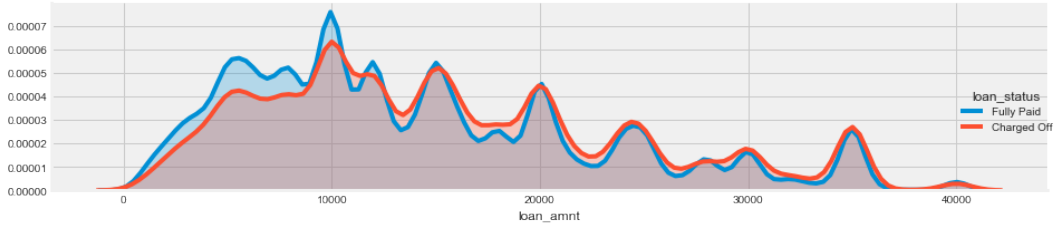
Data Dictionary: "The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value." (Lending Club, 2019)

Figure 17 Loan Amount by Count and Status



Source: own

Figure 18 Loan Amount by Status



Source: own

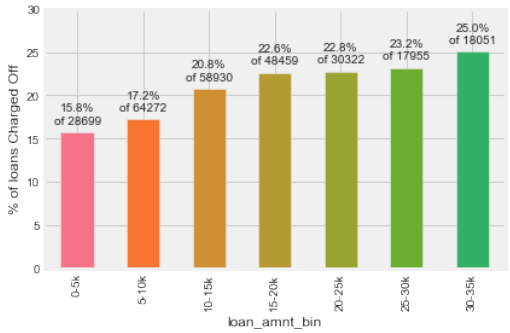
Table 5 Loan Amount by Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	15565.055444	8814.560077	900.0	9000.0	14350.0	20400.0	40000.0
Fully Paid	1076751.0	14134.369808	8669.028973	500.0	7500.0	12000.0	20000.0	40000.0

Source: own

Figure 17, Figure 18 and Table 6 show that Loan Amount(loan_amnt) ranges from 0.5k to 40k, with a median of 12k. "Charged off" loans have a higher loan amount. The Loan Amount(loan_amnt) distribution is slightly right-skewed with mean greater than the median. Most of the loans granted are less than 20 000 (75 %). Around each 5 000 boundary there is a spike in the funding amount. Charged Off loans are shifted when there is a request with a higher average loan amount.

Figure 19 % of Charged Off loans by Loan Amount



Source: own

Figure 19 shows that the % of Charged Off loans increases as we go up the loan amount buckets. The higher is loan amount, even if less loans, the higher risk probability to be Charged Off.

Period(term)

Data Dictionary: "The number of payments on the loan. Values are in months and can be either 36 or 60." (Lending Club, 2019)

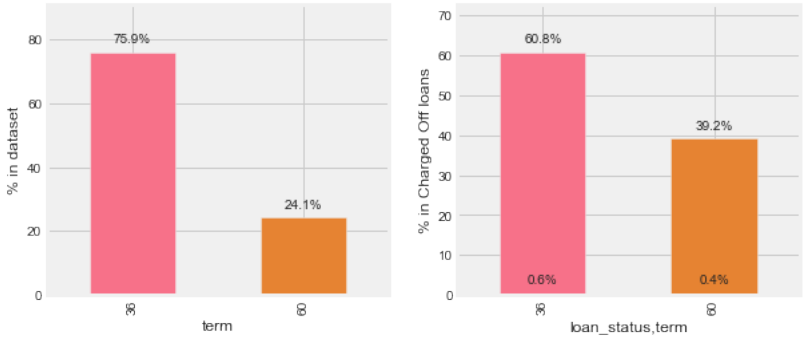
No of loans for each term:

```
36 months    1020743
60 months     324567
```

% of loans for each term:

```
36    0.758742
60    0.241258
```

Figure 20 % of Loans and Charged Off loans by Period(term)



Source: own

Figure 20 shows that about 76% of the loans have the Period(term) of 36 months and the rest 60 months. The 60 months tend to have lower fraction of being Fully Paid. Among Charged Off loans, percentage 60 months gets to 39%.

Charged Off rate compared to loan period Period(term):

```
36    0.159934
60    0.324454
```

16% of the 36 month loans are Charged Off, while 32% of the 60 month are Charged Off which means that an extended term leads to a higher Charged Off rate and this feature is useful in modelling.

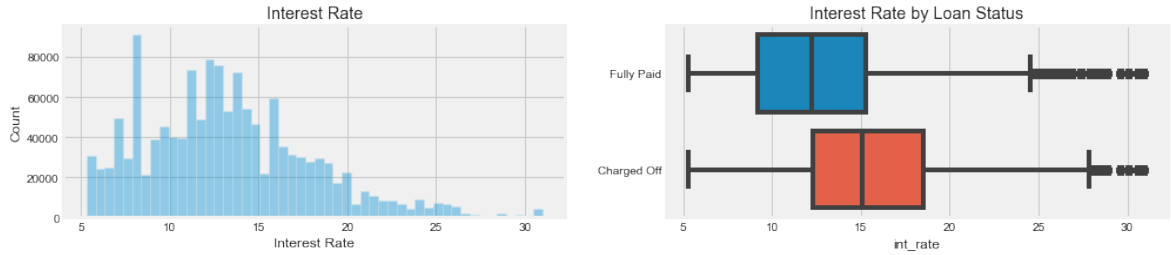
Interest Rate (int_rate)

Data Dictionary: "Interest Rate on the loan."

```
count    1.345310e+06
mean     1.323962e+01
std      4.768716e+00
min      5.310000e+00
25%      9.750000e+00
50%      1.274000e+01
75%      1.599000e+01
max      3.099000e+01
Name: int_rate, dtype: float64
```

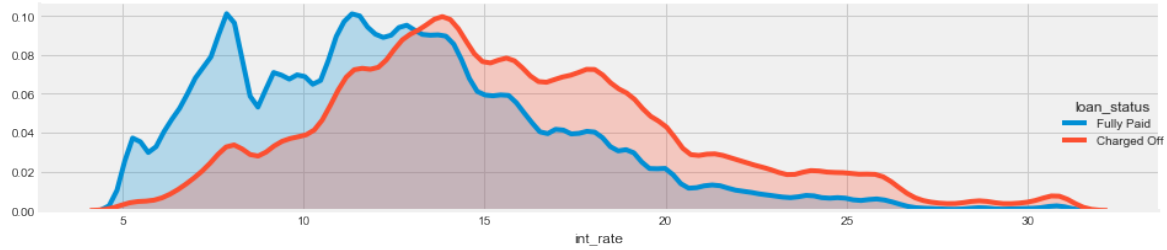
The Interest Rage ranges from 5% to 31%, with a median at 12.7%.

Figure 21 Interest Rate Count by Loan Status



Source: own

Figure 22 Interest Rate by Loan Status



Source: own

Figure 21 shows that the Charged Off loans have higher interest rates and here are quite some outliers for the Fully Paid data instances. Outliers may distort the analysis and should be analysed. For this purpose, the modified Z-score and the IQR method are used. The variable must be continuous, not categorical, for any of these functions to work correctly. The Z-score method is a measure of how many standard deviations below or above the population mean a raw score has. This is a standard score and it could be placed on a normal distribution curve. The IQR method is the interquartile range IQR and is the difference between the 75th and 25th percentile of the dataframe. It is a measure similar to standard variance or deviance, measure of the dispersion, being more robust against outliers.

Table 6 Loan Status by Interest Rate

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	15.710714	4.910387	5.31	12.29	15.05	18.55	30.99
Fully Paid	1076751.0	12.623288	4.527227	5.31	9.17	12.23	15.31	30.99

Source: own

Overall in Table 7 it is shown that the interest rate varies from 5.31% to 30.99% with average interest rate of 15.05%. The interest rate for Charged Off loans is higher than for the Fully Paid. This is quite logic outcome. When the risk increases the rate of interest of the

loan also increases. Any values >75% percentile, 18.5 for "Charged Off" and 15.31 for "Fully Paid" are considered outliers.

Outliers percentage Fully Paid and Charged Off:

```
0.21762598780962358
0.245521468280713
```

Because the percentage of outliers is almost the same for the Fully Paid and Charged Off data, it's too much data to be dropped.

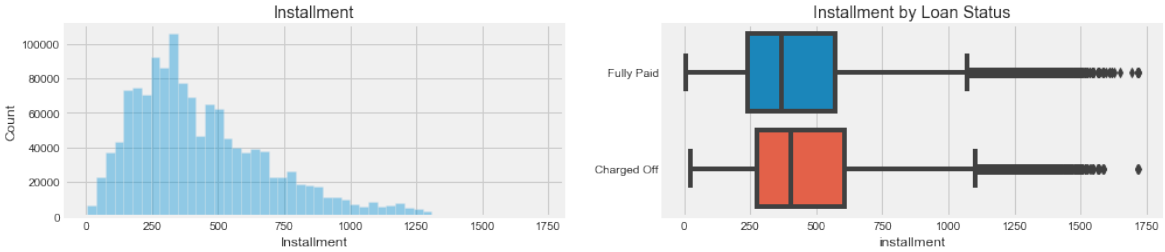
Monthly instalment (instalment)

Data Dictionary: "The monthly payment owed by the borrower if the loan originates." (Lending Club, 2019)

```
count    1.345310e+06
mean     4.380755e+02
std      2.615126e+02
min      4.930000e+00
25%      2.484800e+02
50%      3.754300e+02
75%      5.807300e+02
max      1.719830e+03
Name: installment, dtype: float64
```

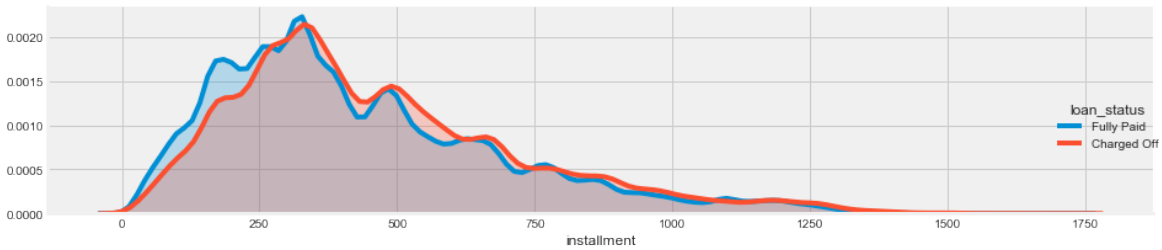
Instalments range from $5 to $1719, with a median at $375.

Figure 23 Instalments by count and Loan Status



Source: own

Figure 24 Instalments by Loan Status



Source: own

From Figure 23 and Figure 24 results that the most common instalment is at about 300. Examining Table 7 can be noticed that the Charged Off loans have higher instalments, but the mean value difference is very small, 275 and 240. In order to see if there are any statistical difference of "instalment" the KS test is done later. The Kolmogorov–Smirnov test is a nonparametric test of the equality of one-dimensional and continuous probability distributions, used to compare a sample with a reference probability distribution or to compare two samples.

Table 7 Installments by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	465.147994	262.774106	21.62	275.29	402.83	605.31	1717.63
Fully Paid	1076751.0	431.323226	260.759563	4.93	240.92	368.26	573.06	1719.83

Source: own

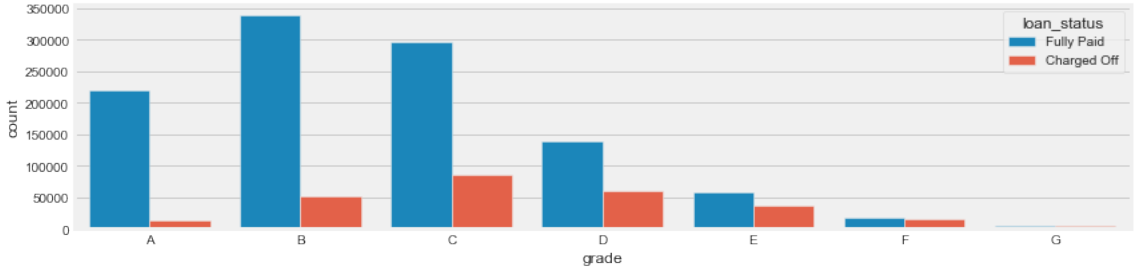
Grade (grade), Sub Grade (sub_grade)

Data Dictionary for grade: "assigned loan grade." Data Dictionary for sub_grade: "LC assigned loan subgrade." (Lending Club, 2019)

Some values of grade and sub_grade:

```
B    392741                          C1    0.063550
C    381686                          B4    0.061844
A    235090                          B5    0.061352
D    200953                          B3    0.060824
E     93650                          C2    0.058881
F     32058
G      9132
```

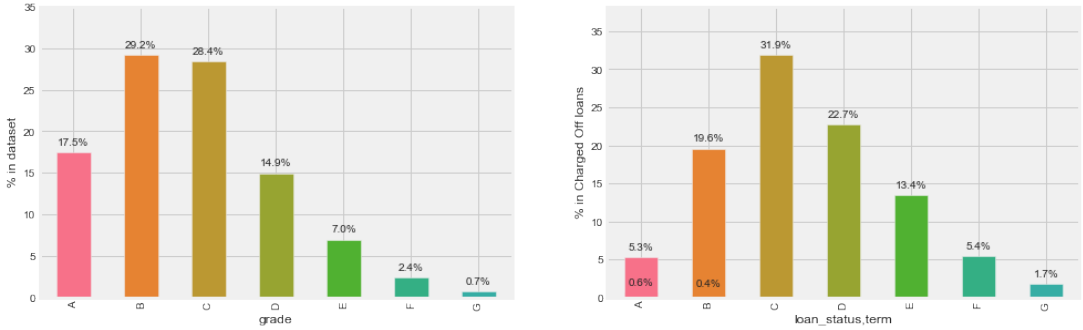
Figure 25 Loans distribution by Grade



Source: own

Figure 25 shows that there is a trend of higher rate of Charged Off loans from grade A to G.

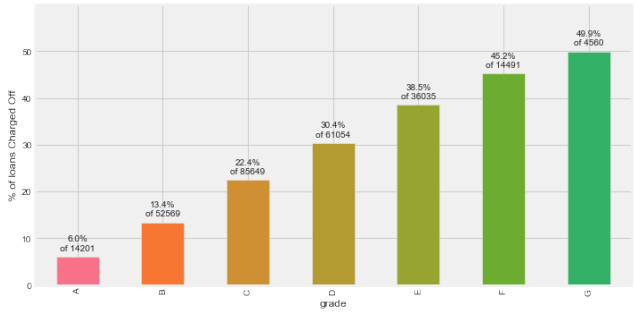
Figure 26 % of Charged off Loans by Grade and Term



Source: own

Figure 26 shows that A grade and B grade loans are the safest. The percentage in the entire dataset is higher than the percentages of the Charged Off loans only. D, E, F, G grade loans are not so safe. Lending Club grading system performs well.

Figure 27 % of Charged off Loans by Grade



Source: own

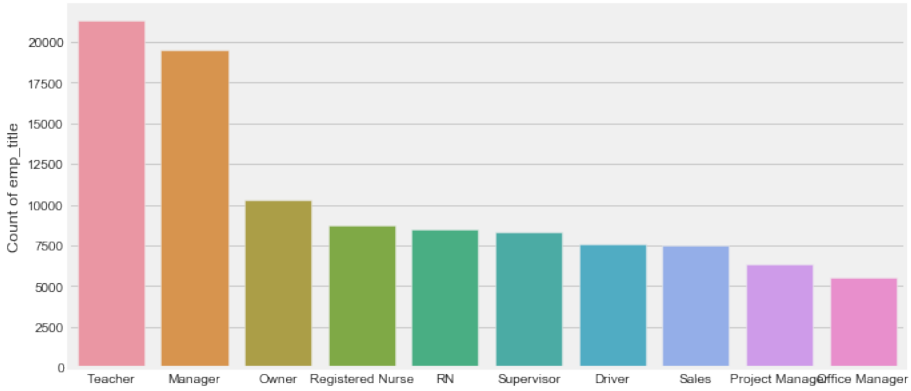
Figure 27 shows that approximately 45% and 49% of all loans in F grade and G grade see are Charged Off. E grade to G grade loans are risky and also less numerous. Lending Club is charging higher interest rate for loans with lower grade, which is correct. The "grade" is contained in the "sub_grade" column, so it’s dropped.

Employment title (emp_title)

Data Dictionary: "The job title supplied by the Borrower when applying for the loan."

```
count      1259525
unique      378353
top        Teacher
freq         21268
```

Figure 28 Employment title



Source: own

Figure 28 shows that maximum of the borrowers are Teachers, then Managers, followed by Owners and Nurses, RN, Supervisors, Drivers, Sales, Project Managers and Office Managers are also among the top 10 Employment Titles (emp_title). There are too many unique names, 378353, in this feature to be considered, so it's dropped.

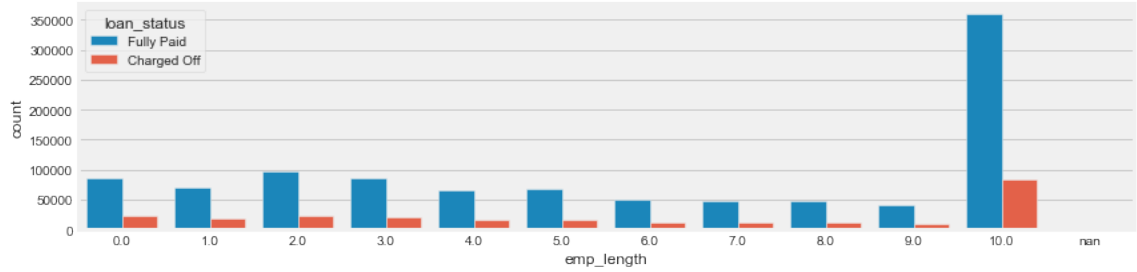
Employment Length (emp_length)

Data Dictionary: "Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years." (Lending Club, 2019) The real data does not match the description:

```
1 year        88494
10+ years    442199
2 years      121743
3 years      107597
4 years       80556
5 years       84154
6 years       62733
7 years       59624
8 years       60701
9 years       50937
< 1 year     108061
NaN           78511
```

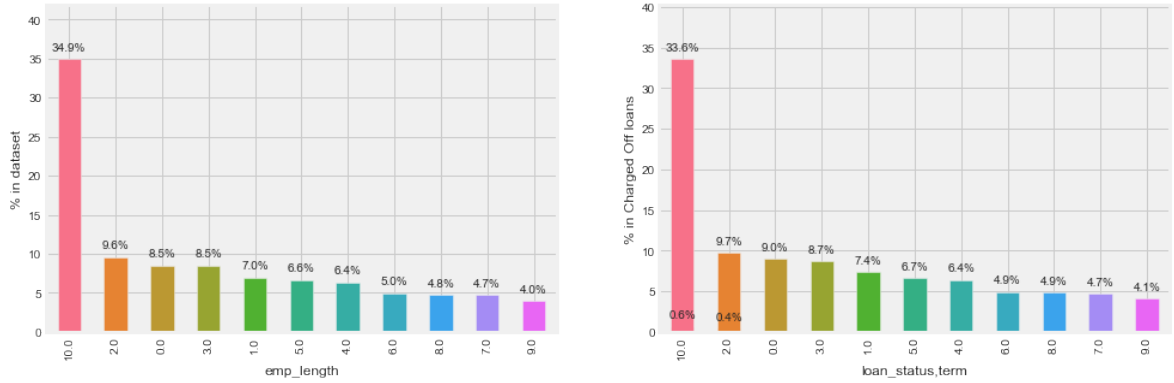
There are 78k loans missing information on Employment Length (emp_length). This feature is converted to integers.

Figure 29 Employment Length by Loan Status



Source: own

Figure 30 Employment Length by %



Source: own

According to the Figure 30, 35% of the of borrowers have an employment length of more than 10 years, which is one third of the loans. The Fully Paid rate doesn't seem to be very related to this feature.

Home Ownerships (home_ownership)

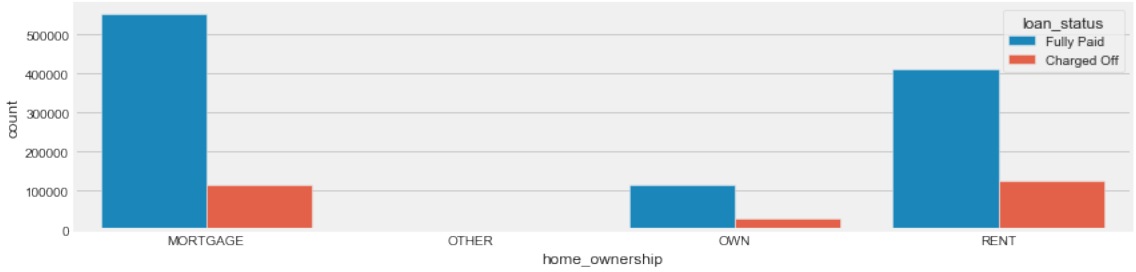
Data Dictionary: "The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER." (Lending Club, 2019)

```
MORTGAGE    665579
RENT        534421
OWN         144832
ANY            286
OTHER          144
NONE            48
```

```
home_ownership
MORTGAGE    0.172077
OTHER       0.188285
OWN         0.206198
RENT        0.232165
```

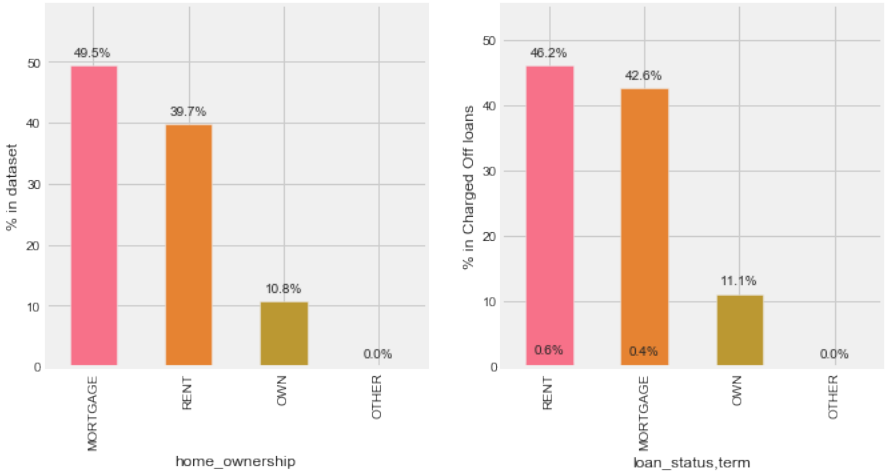
"ANY" and "None" are replaced with "OTHER". From Figure 31 and Figure 32 is visible that there are small differences in Charged Off rates by home ownership. OWN and OTHER have less probability of Charged Off. These differences might be statistically significant because of the large number of instances.

Figure 31 Home Ownership by Loan Status



Source: own

Figure 32 % of Home Ownership



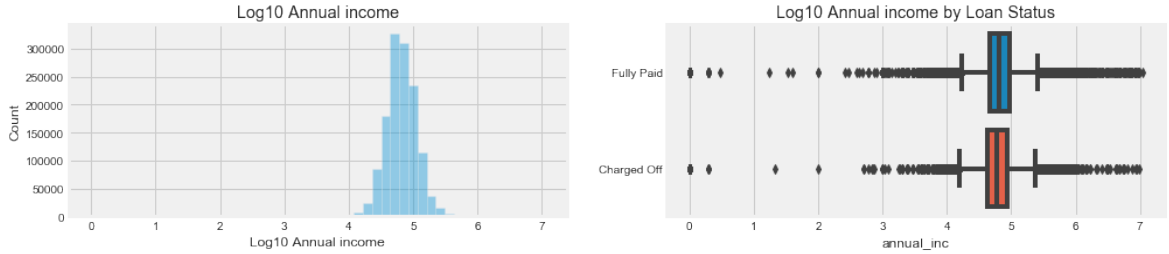
Source: own

Annual Income (annual_inc)

Data Dictionary: "The self-reported annual income provided by the borrower during registration." (Lending Club, 2019)

Because of the large range the values are log-transformed.

Figure 33 Log 10 Annual income



Source: own

Table 8 Annual Income by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	4.781756	0.241984	0.0	4.633479	4.778158	4.929424	6.977724
Fully Paid	1076751.0	4.820560	0.248526	0.0	4.672107	4.812920	4.968488	7.041361

Source: own

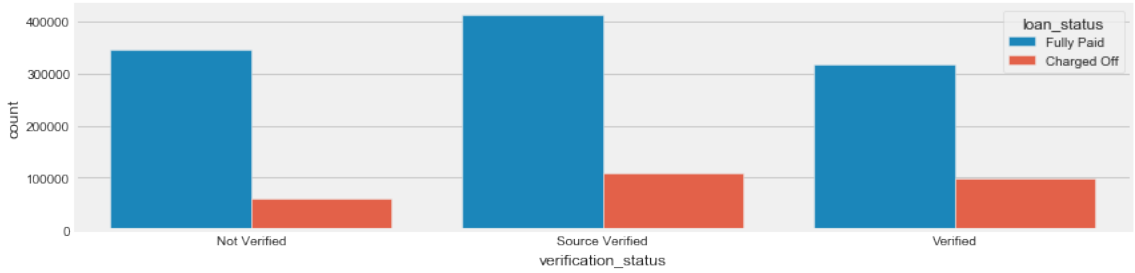
Examining Table 8 and the Figure 10, for both cases, there are quite some outliers. The annual income ranges from min of 0 to max of 7,000,000 with the median around 48,000. Most people have an annual income less than 4,633,00.

Verification Status (verication_status)

Data Dictionary: "Indicates if income was verified, not verified, or if the income source was verified by Lending Club.“

```
Source Verified    521273
Verified           418336
Not Verified       405701
Name: verification_status, dtype: int64
```

Figure 34 Verification Status by Loan Status



Source: own

Unexpectedly, Figure 34 shows that verified loans have a higher chance to be Charged Off. The “Verified” and “Source Verified” loan status mean that Lending Club has cross checked the income value specified by the borrower and because of this both statuses can be merged.

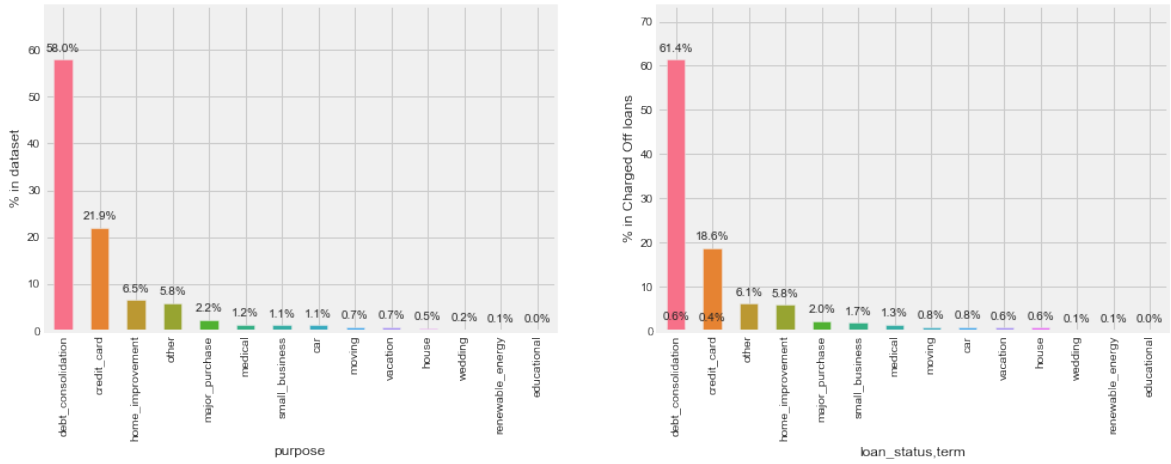
Purpose (purpose)

Data Dictionary: "A category provided by the borrower for the loan request." (Lending Club, 2019)

```
debt_consolidation    780321
credit_card           295279
home_improvement       87504
other                  77875
major_purchase         29425
medical                15554
small_business         15416
car                    14585
moving                  9480
vacation                9065
house                   7253
wedding                 2294
renewable_energy         933
educational              326
Name: purpose, dtype: int64
```

The Charged Off rates of purposes are the most interesting.

Figure 35 % of Loans by Purpose



Source: own

Figure 35 shows that 58% of the loans have the purpose of dept consolidation. 18.6% of the loans with debit_card purpose are Charged Off, but the number is not too big to be significant.

Title (title)

Data Dictionary: "The loan title provided by the borrower." (Lending Club, 2019)

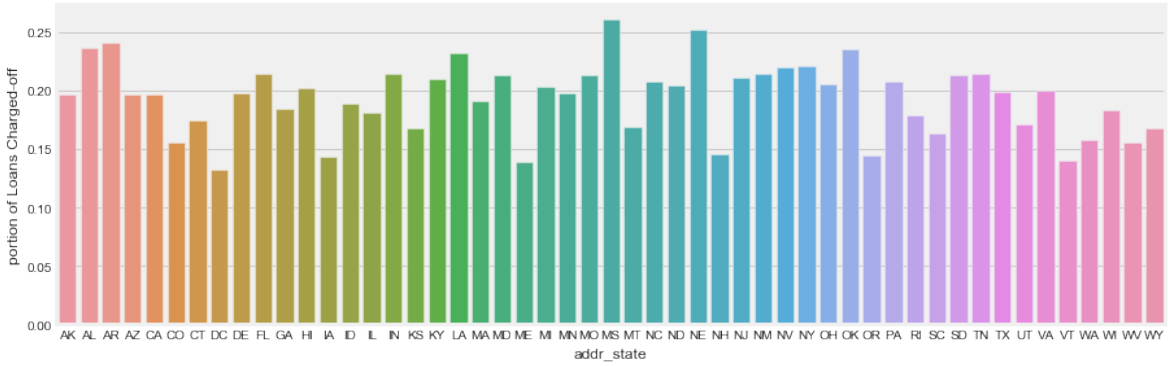
```
count                 1328651
unique                  61682
top       Debt consolidation
freq                   660960
Name: title, dtype: object

Debt consolidation          660960
Credit card refinancing     249910
Home improvement             75073
Other                        66778
Major purchase               24045
Debt Consolidation           15695
Medical expenses             13576
Business                     11750
Car financing                11220
Vacation                      8046
Name: title, dtype: int64
```

Title column has 61682 unique values describing the loan. This information is already included in purpose column under 14 categories. It's dropped.

Zip Code (zip_code), State (addr_stat)

Data Dictionary for zip_code: "The first 3 numbers of the zip code provided by the borrower in the loan application." Data Dictionary for addr_state: "The state provided by the borrower in the loan application.

Figure 36 % of Charged Off Loans



Source: own

Figure 36 shows that MS has the highest rate of Charged Off, 0.26, and DC has the lowest rate, 0.13. The percentage is not considered significant and is to be ignored. There

are many unique "zip_code" values, and the zip_code info should be equivalent to "addr_stat". The "zip_code" is dropped.

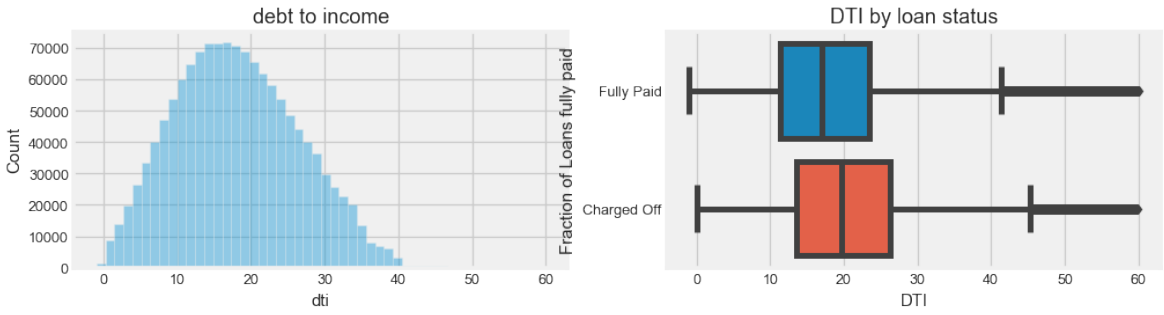
Debt to Income (dti)

Data Dictionary: "A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested Lending Club loan, divided by the borrower's self-reported monthly income." (Lending Club, 2019)

```
count    1.344936e+06
mean     1.828267e+01
std      1.116045e+01
min     -1.000000e+00
25%      1.179000e+01
50%      1.761000e+01
75%      2.406000e+01
max      9.990000e+02
Name: dti, dtype: float64
```

The values of -1 or 999 are probably outliers, which can be seen in the histogram from Figure 37.

Figure 37 DTI count and outliers



Source: own

Table 9 DTI by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268488.0	20.171196	11.831747	0.0	13.55	19.76	26.31	999.0
Fully Paid	1076448.0	17.811629	10.935927	-1.0	11.43	17.11	23.44	999.0

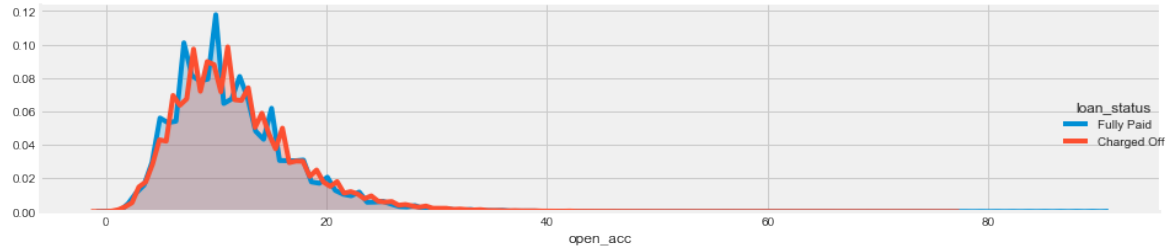
Source: own

Examining the Table 9 can be noticed that when the Debt to income (dti) ratio is higher than 20%, higher percentage of loans are being Charged Off. If the Debt to income (dti) is higher the chances of loan to be Charged Off are also higher.

No of Open Credit Lines (open_acc)

Data Dictionary: "The number of open credit lines in the borrower's credit file." (Lending Club, 2019)

Figure 38 No of credit lines by Loan Status



Source: own

Figure 38 shows that there is no significant difference between Fully Paid loans and Charged Off loans in no of credit lines.

Table 10 No of Credit Lines by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	11.901269	5.637964	0.0	8.0	11.0	15.0	76.0
Fully Paid	1076751.0	11.516763	5.429354	0.0	8.0	11.0	14.0	90.0

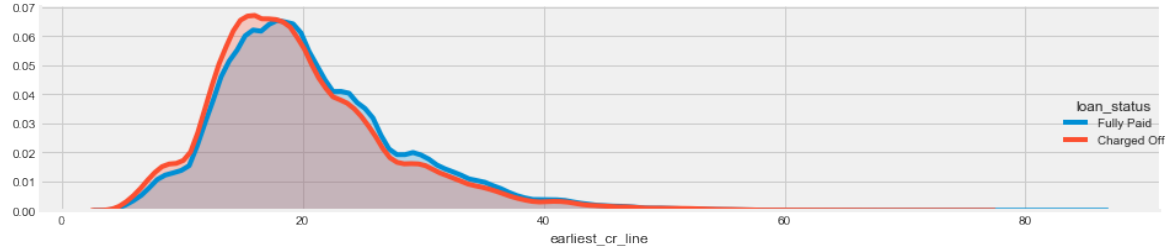
Source: own

In Table 10 can be seen from the median of the distribution that maximum the borrowers have 15 open credit lines. There is no significant difference in the credit lines of the Fully Paid and Charged Off loans.

Earliest Credit Line month (earliest_cr_line)

Data Dictionary: "The month the borrower's earliest reported credit line was opened." (Lending Club, 2019)

Figure 39 Earliest Credit Line by Loan Status



Source: own

Charged Off borrowers have opened their first credit line more recently, but as we see in Table 11 the medians are almost the same.

Table 11 Earliest Credit Line by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	19.893397	7.583253	4.142466	14.813699	18.569863	23.652055	75.772603
Fully Paid	1076751.0	20.731599	7.613510	3.975342	15.482192	19.317808	24.652055	85.531507

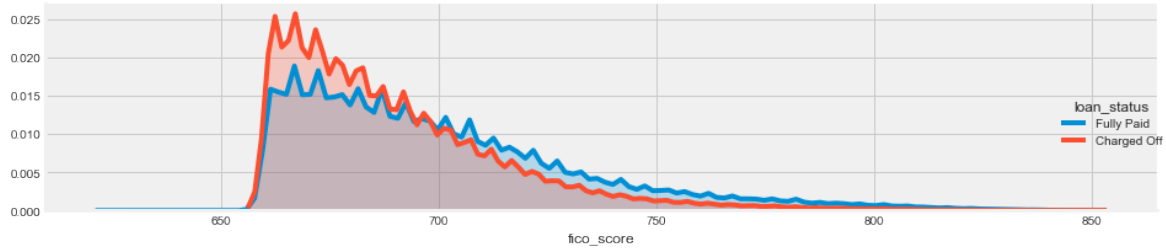
Source: own

FICO (fico_range_low, fico_range_high)

Data Dictionary for fico_range_low: "The lower boundary range the borrower's FICO at loan origination belongs to." (Lending Club, 2019) Data Dictionary for fico_range_high: "The upper boundary range the borrower's FICO at loan origination belongs to." (Lending Club, 2019)

"fico_range_low" and "fico_range_high" are highly correlated, but one new feature will be created which called "fico_score" and is the mean value of the two initial values.

Figure 40 FICO Score by Loan Status



Source: own

Figure 40 shows that there is quite some difference in FICO scores between Fully Paid and Charged Off loans.

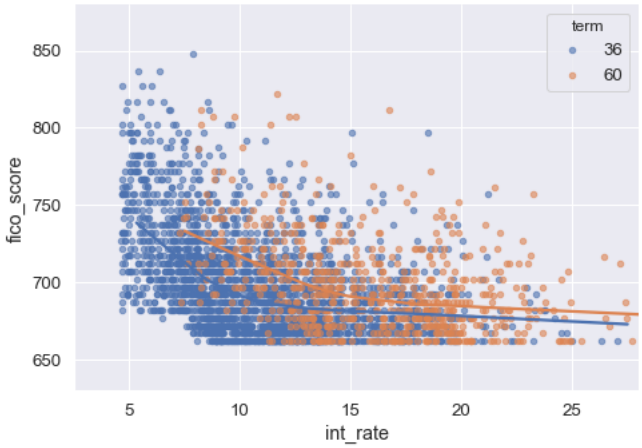
Because FICO score is an important feature, the correlations(Table 12) between FICO score, int_rate and the term are checked as these are 3 of the most important features that have different distributions between Charged Off and Fully Paid.

Table 12 Correlations between FICO Score and other 3 features

	fico_score	int_rate	term
fico_score	1.000000	-0.405365	-0.002222
int_rate	-0.405365	1.000000	0.417230
term	-0.002222	0.417230	1.000000

Source: own

Figure 41 FICO Score and Interest Rate Correlation



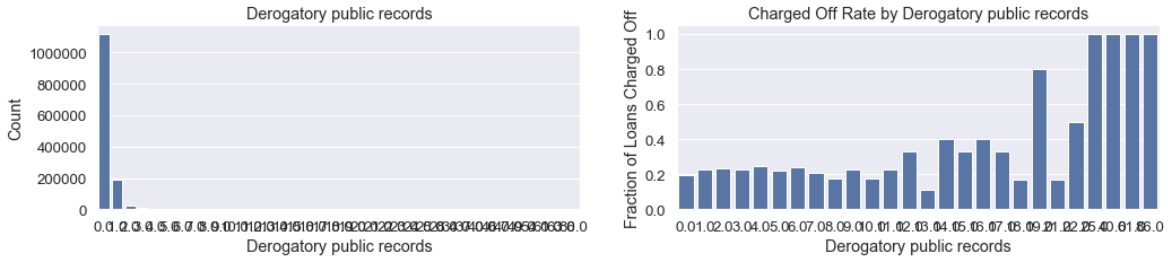
Source: own

Figure 41 shows that the higher the fico_score is, the lower is the int_rate. 36 month loans have higher fico_score at lower int_rate.

Derogatory Public Records (pub_rec)

Data Dictionary: "Number of derogatory public records." (Lending Club, 2019)

Figure 42 Derogatory Public Records



Source: own

Figure 42 shows that Derogatory public records (pub_rec) > 11 have a significant Charged Off rate but it's affected by low counts. Considering the Table 13 overall statistics, the mean difference in Derogatory public records (pub_rec) is just 0.2 for Fully Paid and Charged Off.

Table 13 Derogatory Public Records by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	0.246843	0.656732	0.0	0.0	0.0	0.0	86.0
Fully Paid	1076751.0	0.207403	0.587118	0.0	0.0	0.0	0.0	63.0

Source: own

Few bins with low values are grouped into 1 bin, Table 13.

Table 14 Binary Derogatory Public Records by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	0.236347	0.537366	0.0	0.0	0.0	0.0	3.0
Fully Paid	1076751.0	0.198903	0.498270	0.0	0.0	0.0	0.0	3.0

Source: own

From Table 14 is clear that Derogatory public records (pub_rec) are mostly 0 and 1; Derogatory public records (pub_rec) with >1 have higher chance to be Charged Off than Derogatory public records (pub_rec) of 0. 94% of borrowers have no public records and 5% have 1. Even if there is 1 derogatory record the chances of Charged Off increase significantly.

Public Record Bankruptcies (pub_rec_bankruptcies)

Data Dictionary: "Number of public record bankruptcies." (Lending Club, 2019)

Table 15 Number of public record and public record bankruptcies

	pub_rec	pub_rec_bankruptcies
pub_rec	1.000000	0.735344
pub_rec_bankruptcies	0.735344	1.000000

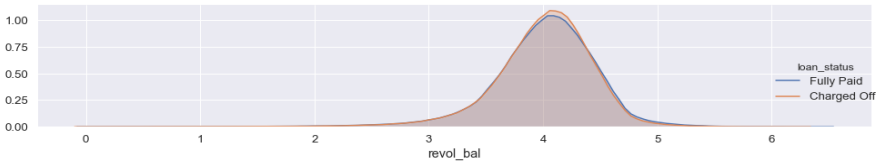
Source: own

From Table 15 is visible that Derogatory public records (pub_rec) and Public record bankruptcies (pub_rec_bankruptcies) are highly correlated. Only one of them can be used in the modelling.

Revolving Balance (revol_bal)

Data Dictionary: "Total credit revolving balance." (Lending Club, 2019)

Figure 43 Credit revolving balance by Loan Status



Source: own

From Figure 43 is visible that there is no big difference in revolving balance mean for Charged Off and Fully Paid loans.

Revolving Line Utilization Rate (revol_util)

Data Dictionary: "Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit." (Lending Club, 2019)

Table 16 Revolving line utilization rate

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268379.0	54.758433	23.865982	0.0	37.4	55.5	73.1	366.6
Fully Paid	1076074.0	51.074671	24.626821	0.0	32.5	51.3	70.1	892.3

Source: own

Table 16 shows that Charged Off loans have higher Revolving Line Utilization Rate (revol_util). The average Revolving Line Utilization Rate (revol_util) rate is around 54% .

Total no of Credit Lines (total_acc)

Data Dictionary: "The total number of credit lines currently in the borrower's credit file." (Lending Club, 2019)

Table 17 No of Credit Lines by Loan Status

loan_status	count	mean	std	min	25%	50%	75%	max
Charged Off	268559.0	24.709356	12.117364	2.0	16.0	23.0	31.0	176.0
Fully Paid	1076751.0	25.048551	11.967683	2.0	16.0	23.0	32.0	162.0

Source: own

The are approximately 23 average credit lines for the borrower's. There is no big difference between the Charged Off and Fully Paid credit lines (Table 17).

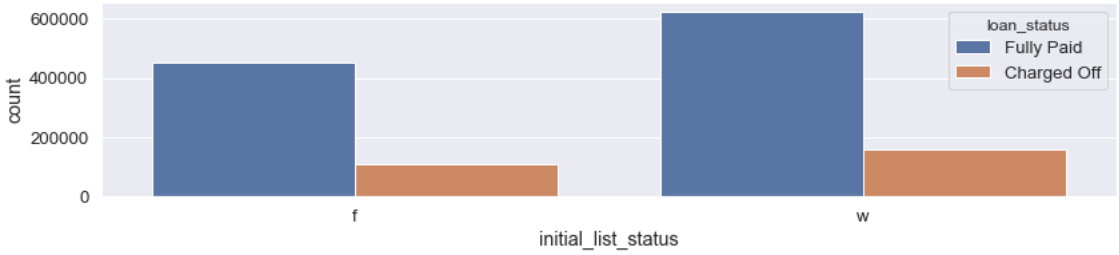
Initial Listing Status (initial_list_status)

Data Dictionary: "The initial listing status of the loan. Possible values are – W, F (whole vs. fractional)“ (Lending Club, 2019)

```
w    784010
f    561300
Name: initial_list_status, dtype: int64
```

More granted loans have whole amount and the rest have fractional amount. Figure 44 shows that "Whole" has higher rate of Charged-Off.

Figure 44 Whole vs. Fractional Loans



Source:own

Publicity Availability (policy_code)

Data Dictionary: "publicly available policy_code=1; new products not publicly available policy_code=2" (Lending Club, 2019)

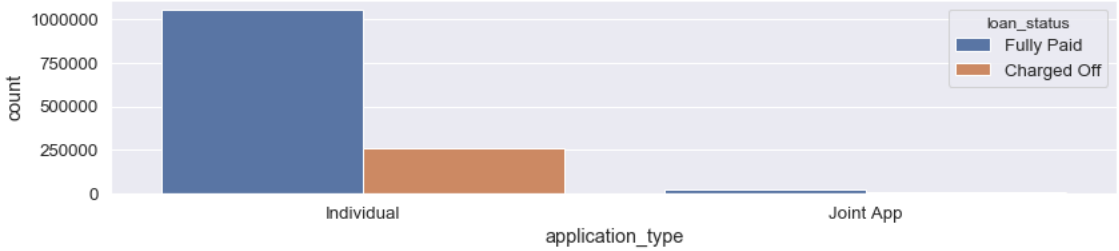
```
1.0    1345310
Name: policy_code, dtype: int64
```

Application type (application_type)

Data Dictionary: "Indicates whether the loan is an individual application or a joint application with two co-borrowers." (Lending Club, 2019)

```
Individual    1319510
Joint App       25800
Name: application_type, dtype: int64
```

Figure 45 Application Type by Loan Status



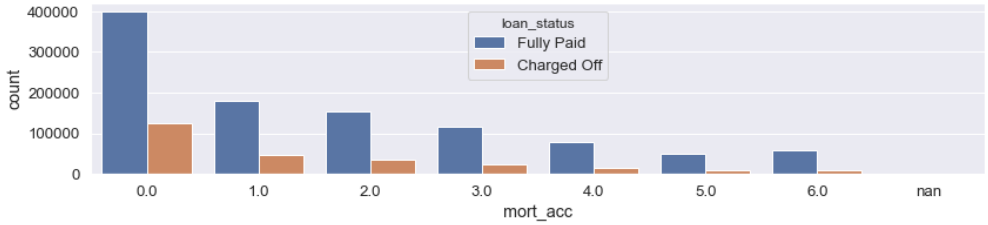
Source: own

Figure 45 shows that most of the loans are individual applications and these are also more probable to be Charged Off.

No of Mortgage Accounts (mort_acc)

Data Dictionary: "Number of mortgage accounts." (Lending Club, 2019)

Figure 46 No of Mortgage Accounts by Loan Status



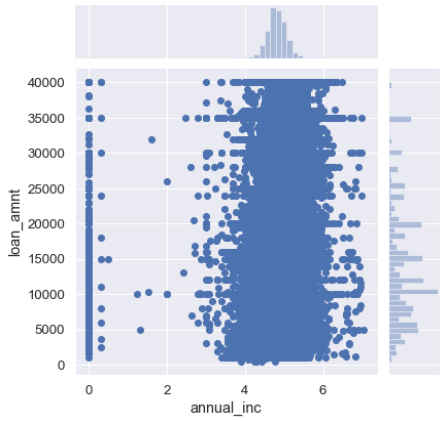
Source: own

Figure 46 shows that if the borrower has more mortgage accounts, the lower chances of loans to be Charged Off.

4.3 Bivariate analysis

Loan Amount (loan_amnt) and Annual Income (annual_inc)

Figure 47 Loan Amount and Annual Income

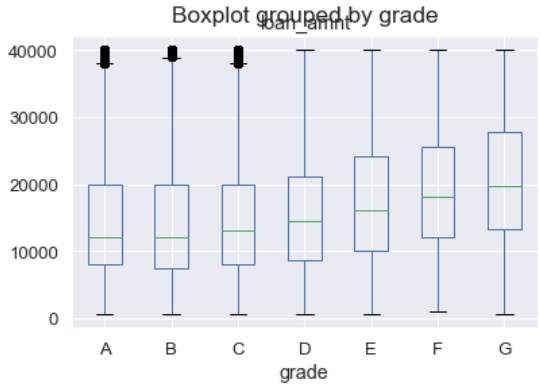


Source: own

Figure 47 shows that there are borrowers with an average income which is lower than 50 000 taking loans of 25 000 or higher which are risky loans.

Loan Amount (loan_amnt) and Grade (grade)

Figure 48 Loan Amount and Grade

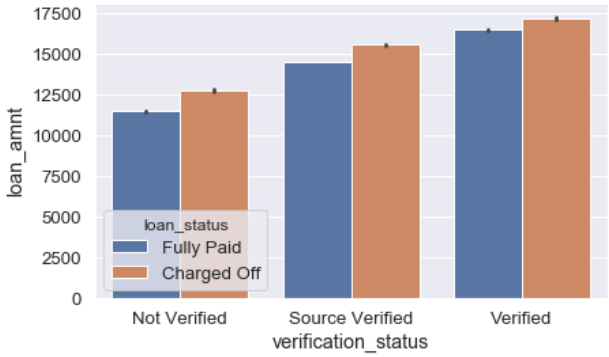


Source: own

Figure 48 shows that lower grade is given to a larger loans, and the median amount for G grade which is almost 10 000 higher than A, B, or C grade loans.

Loan Amount (loan_amnt) and Verification Status (verification_status) and Loan Status(loan_status)

Figure 49 Loan Amount and Verification Status and Loan Status

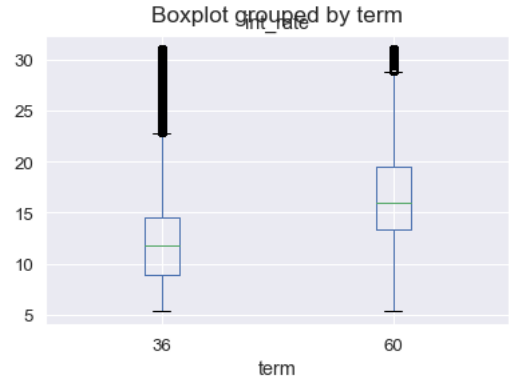


Source: own

Figure 49 shows that more often are Verified loan with higher amount. Is known that there are less loans with a high amount, but they have a high Charge Off rate. It's not the verified status per se, it's the fact that higher loan amounts are riskier and are also verified more often by Lending Club.

Interest Rate(int_rate) and Period(term)

Figure 50 Interest Rate and Term

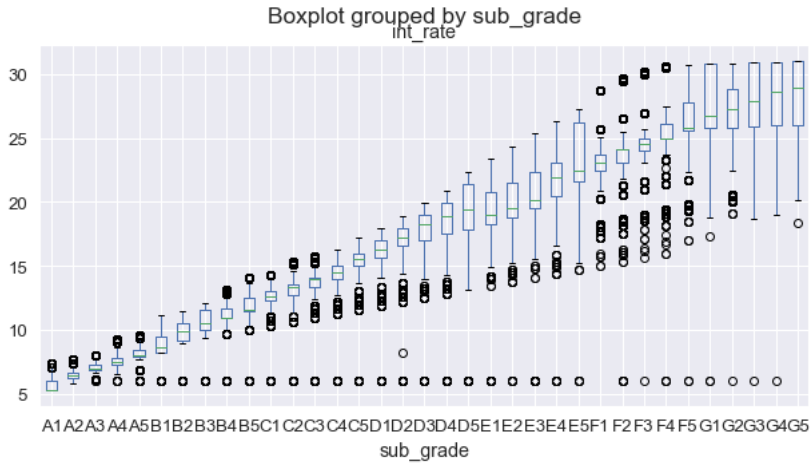


Source: own

Interest rates are connected with the term. Higher amounts have a higher term. In Figure 50 is shown that the interest rate associated with them is also high.

Interest Rate(int_rate) and Sub Grade(sub_grade)

Figure 51 Interest Rate and Sub Grade

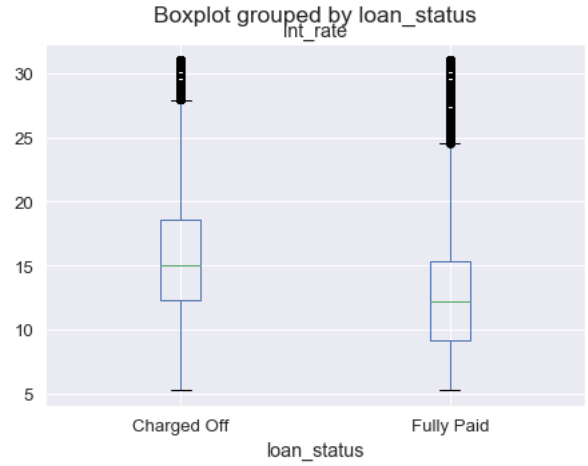


Source: own

Figure 51 shows how the interest rate varies quite directly with the subgrade. Lower the sub grade, result in higher the interest.

Interest Rate(int_rate) and Loan Status (loan_status)

Figure 52 Interest Rate and Loan Status

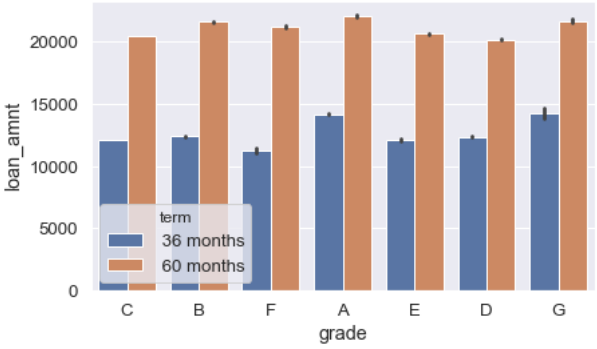


Source: own

Figure 52 shows that loans with a higher interest rate are more probable to be Charged Off.

Loan Amount(loan_amnt) and Grade(grade) and Period(term)

Figure 53 Loan Amount and Grade and Term

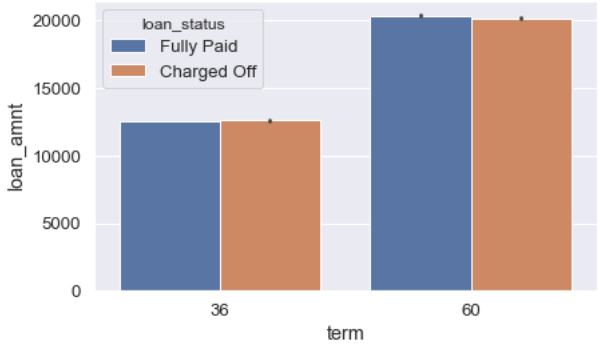


Source: own

The assumptions from the univariate analysis are more obvious in this plot(Figure 53). If the loan amount is higher then the grade for long term is lower.

Loan Amount (loan_amnt) and Period(term) and Loan Status(loan_status)

Figure 54 Loan Amount and Term and Loan Status

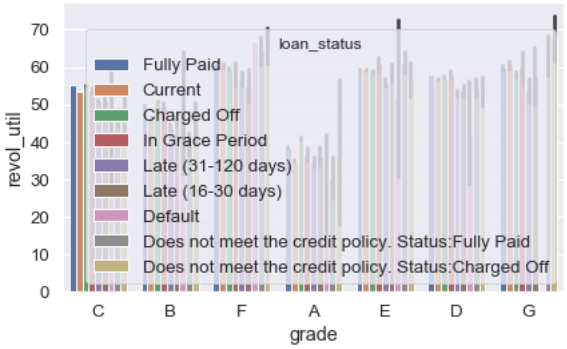


Source: own

The assumptions from the univariate analysis are more obvious in this plot (Figure 54). If the loan amount is higher than the long term occurs more often to be Charged Off.

Revolving Line Utilization Rate (revol_util) and Grade(grade) and Loan Status(loan_status)

Figure 55 Revol_util and Grade and Loan Status



Source: own

Figure 55 shows that Revolving Line Utilization Rate (revol_util) and grade are correlated. The Revolving Line Utilization Rate (revol_util) is positivly correlated to the grade because the grade begins from A to E and the Revolving Line Utilization Rate (revol_util) increases same direction. It can happen because higher loan amounts get higher grades.

5 Results and Discussion

5.1 Model Building

All the remaining features are considered individually in this part of the thesis. Overall statistical tests are performed for all the features on distributions and correlations. Also, the target variable is being converted to 0 and 1 indicator because converting the target variable to numerical variable is convenient and useful for the prediction scope. Fully Paid: 0, Charged Off :1.

In this section the loans are being grouped to "Fully Paid" and "Charged Off", and then hypothesis tests are used to compare the distributions of each feature. If the test statistic is small or the p-value is high (>0.05, 95% confidence level), the null hypothesis is accepted and means that the distributions of the two samples are the same, but if p<0.05 it means that the distributions are different.

K-S Tests, Z Tests and Chi Squared Tests:

The results of K-S tests:

```
loan_amnt:
Ks_2sampResult(statistic=0.08341793565761396, pvalue=0.0)
term:
Ks_2sampResult(statistic=0.1884875690616944, pvalue=0.0)
int_rate:
Ks_2sampResult(statistic=0.26709123944420854, pvalue=0.0)
installment:
Ks_2sampResult(statistic=0.07177308649462699, pvalue=0.0)
emp_length:
Ks_2sampResult(statistic=0.02538830065622577, pvalue=9.11735879570968e-121)
annual_inc:
Ks_2sampResult(statistic=0.07162968406367032, pvalue=0.0)
dti:
Ks_2sampResult(statistic=0.11413770318986116, pvalue=0.0)
earliest_cr_line:
Ks_2sampResult(statistic=0.05478115417211504, pvalue=0.0)
open_acc:
Ks_2sampResult(statistic=0.02923793065231428, pvalue=4.9928027835478835e-160)
pub_rec:
Ks_2sampResult(statistic=0.029598859981289527, pvalue=5.414690789546113e-164)
revol_bal:
Ks_2sampResult(statistic=0.018292907328583796, pvalue=6.685708285584415e-63)
revol_util:
Ks_2sampResult(statistic=0.06461830322886436, pvalue=0.0)
total_acc:
Ks_2sampResult(statistic=0.017750078823176307, pvalue=3.0059713277050153e-59)
mo_sin_old_il_acct:
Ks_2sampResult(statistic=0.03536030133651713, pvalue=7.229532079723465e-234)
mo_sin_old_rev_tl_op:
Ks_2sampResult(statistic=0.06674118284428288, pvalue=0.0)
mort_acc:
Ks_2sampResult(statistic=0.09312970723716452, pvalue=0.0)
pub_rec_bankruptcies:
Ks_2sampResult(statistic=0.020872747616026865, pvalue=9.133854484547676e-82)
```

```
fico_score:
Ks_2sampResult(statistic=0.1386761979596195, pvalue=0.0)
Charged_Off:
Ks_2sampResult(statistic=1.0, pvalue=0.0)
```

Z tests results:

```
loan_amnt:
Ks_2sampResult(statistic=0.08341793565761396, pvalue=0.0)
term:
Ks_2sampResult(statistic=0.1884875690616944, pvalue=0.0)
int_rate:
Ks_2sampResult(statistic=0.26709123944420854, pvalue=0.0)
installment:
Ks_2sampResult(statistic=0.07177308649462699, pvalue=0.0)
emp_length:
Ks_2sampResult(statistic=0.02538830065622577, pvalue=9.11735879570968e-121)
annual_inc:
Ks_2sampResult(statistic=0.07162968406367032, pvalue=0.0)
dti:
Ks_2sampResult(statistic=0.11413770318986116, pvalue=0.0)
earliest_cr_line:
Ks_2sampResult(statistic=0.05478115417211504, pvalue=0.0)
open_acc:
Ks_2sampResult(statistic=0.02923793065231428, pvalue=4.9928027835478835e-160)
pub_rec:
Ks_2sampResult(statistic=0.029598859981289527, pvalue=5.414690789546113e-164)
revol_bal:
Ks_2sampResult(statistic=0.018292907328583796, pvalue=6.685708285584415e-63)
revol_util:
Ks_2sampResult(statistic=0.06461830322886436, pvalue=0.0)
total_acc:
Ks_2sampResult(statistic=0.017750078823176307, pvalue=3.0059713277050153e-59)
mo_sin_old_il_acct:
Ks_2sampResult(statistic=0.03536030133651713, pvalue=7.229532079723465e-234)
mo_sin_old_rev_tl_op:
Ks_2sampResult(statistic=0.06674118284428288, pvalue=0.0)
mort_acc:
Ks_2sampResult(statistic=0.09312970723716452, pvalue=0.0)
pub_rec_bankruptcies:
Ks_2sampResult(statistic=0.020872747616026865, pvalue=9.133854484547676e-82)
fico_score:
Ks_2sampResult(statistic=0.1386761979596195, pvalue=0.0)
Charged_Off:
Ks_2sampResult(statistic=1.0, pvalue=0.0)
```

Chi-squared Tests results:

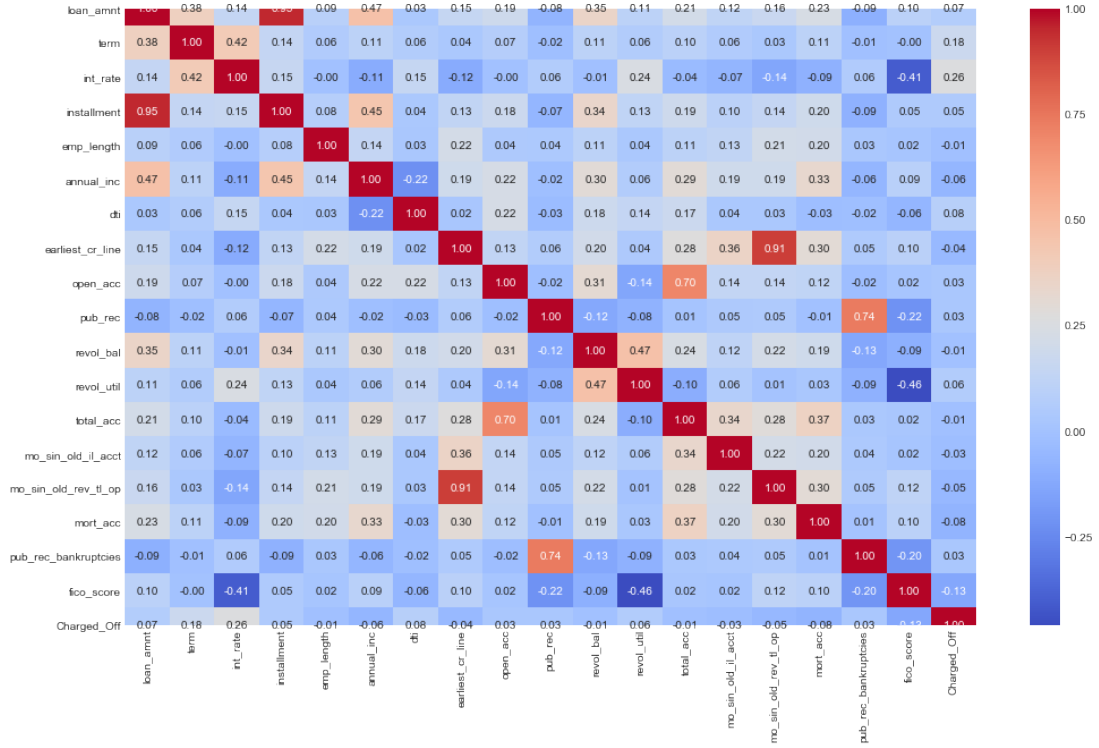
```
loan_amnt:
Ks_2sampResult(statistic=0.08341793565761396, pvalue=0.0)
term:
Ks_2sampResult(statistic=0.1884875690616944, pvalue=0.0)
int_rate:
Ks_2sampResult(statistic=0.26709123944420854, pvalue=0.0)
installment:
Ks_2sampResult(statistic=0.07177308649462699, pvalue=0.0)
emp_length:
Ks_2sampResult(statistic=0.02538830065622577, pvalue=9.11735879570968e-121)
annual_inc:
Ks_2sampResult(statistic=0.07162968406367032, pvalue=0.0)
dti:
Ks_2sampResult(statistic=0.11413770318986116, pvalue=0.0)
earliest_cr_line:
Ks_2sampResult(statistic=0.05478115417211504, pvalue=0.0)
open_acc:
Ks_2sampResult(statistic=0.02923793065231428, pvalue=4.9928027835478835e-160)
pub_rec:
```

```
Ks_2sampResult(statistic=0.029598859981289527, pvalue=5.414690789546113e-164)
revol_bal:
Ks_2sampResult(statistic=0.018292907328583796, pvalue=6.685708285584415e-63)
revol_util:
Ks_2sampResult(statistic=0.06461830322886436, pvalue=0.0)
total_acc:
Ks_2sampResult(statistic=0.017750078823176307, pvalue=3.0059713277050153e-59)
mo_sin_old_il_acct:
Ks_2sampResult(statistic=0.03536030133651713, pvalue=7.229532079723465e-234)
mo_sin_old_rev_tl_op:
Ks_2sampResult(statistic=0.06674118284428288, pvalue=0.0)
mort_acc:
Ks_2sampResult(statistic=0.09312970723716452, pvalue=0.0)
pub_rec_bankruptcies:
Ks_2sampResult(statistic=0.020872747616026865, pvalue=9.133854484547676e-82)
fico_score:
Ks_2sampResult(statistic=0.1386761979596195, pvalue=0.0)
Charged_Off:
Ks_2sampResult(statistic=1.0, pvalue=0.0)
```

According to the results of the K-S Test, Z Tests and Chi Square Tests, the null hypothesis is accepted for the emp_length, open_acc, pub_rec, revol_bal, total_acc, mo_sin_old_il_acct, pub_rec_bankruptcies which means that the distributions of these variables are the same with the loan_status. The remaining variables have different distributions. But, all the features have different distributions for the Loan Status(Charged_off).

Pearson Correlation

Figure 56 Pearson Correlation



Source: own

According to the Figure 56 the linearly correlated features are:

- "installment" and "loan_amnt": 0.95
- "mo_sin_old_rev_tl_op"* and "earliest_cr_line": 0.91
- "pub_rec_bankruptcies"* and "pub_rec": 0.74
- "total_acc" and "open_acc": 0.70

(*) has null values

Pearson Correlation, Figure 56, was done with the scope to identify the highly correlated features and potentially remove them as they could impact the outcome of the models but none of them is removed in the end since they are playing an important role in the lending process.

Table 18 Pearson Correlation Results sorted by abs val

	index	pearson_corr
0	int_rate	0.258792
1	term	0.176096
2	fico_score	-0.130683
3	dti	0.084510
4	mort_acc	-0.077832
5	loan_amnt	0.065604
6	annual_inc	-0.062614
7	revol_util	0.060048
8	installment	0.051701
9	mo_sin_old_rev_tl_op	-0.050502
10	earliest_cr_line	-0.043999
11	pub_rec	0.029548
12	open_acc	0.028078
13	mo_sin_old_il_acct	-0.026213
14	pub_rec_bankruptcies	0.025308
15	emp_length	-0.014235
16	total_acc	-0.011300
17	revol_bal	-0.005929

Source: own

According to the Table 18, the features which are most linearly correlated with the target variable(Charged_Off) are interest rate, loan term, Fico Score and debt-to-income ratio. The least correlated features are the revolving balance, employment length, and total_acc. The dependence of the Charged Off on the predictors shows that the "int_rate" is the most correlated one.

5.1.1 Feature Engineering

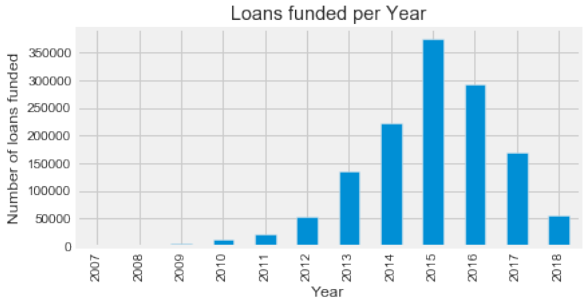
In this section the remaining data is cleaned and prepared for machine learning. The Feature Engineering includes One-Hot Encoding, training and test split, imputing missing values, Scaling, PCA, Linear Dependence of the Charged Off.

Training and Test data split

The loans funded on earlier dates are used to predict the outcome of loans funded on later dates. There are 139 unique values. The first loans were issued in Jan 2007 and the last loans were issued in December 2018. The busiest month was March 2016, with 48937 loans funded.

```
count                 1345310
unique                    139
top       2016-03-01 00:00:00
freq                    48937
first     2007-06-01 00:00:00
last      2018-12-01 00:00:00
```

Figure 57 Loans Funded per Year



Source: own

Figure 57 shows that the funded loans increased significantly from 2013, and continued till 2015. The train and test sample is split at 7:3.

Imputing missing values

An imputer object will be created with the strategy set to the median. This object is trained on the training data and used to fill in the missing values in both the test and training data. The missing values in test data are filled in with the median value from training data. This is done to avoid the test data leakage problem. Table 19 contains all the features which have missing values.

Table 19 Features with Missing Values

	Missing Values	% of Total Values	type
mo_sin_old_il_acct	105575	7.8	float64
emp_length	78511	5.8	float64
mort_acc	47281	3.5	float64
revol_util	857	0.1	float64
dti	374	0.0	float64

Source: own

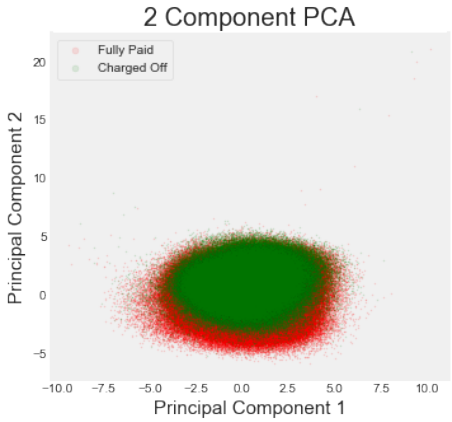
Feature Scaling

As features are measured using different units, they cover different ranges. Some Machine Learning models, such as SVM, KNN that take into account distance measures between observations are therefore significantly affected by the range of the features and scaling allows them to learn. While some methods, such as Linear Regression and Random Forest do not actually require feature scaling, it's still best practice to take this step when comparing multiple algorithms.

PCA visualization of the train data

PCA analysis is done to check if some combinations of variables that explain the most variance can be found. To visualize the dataset a 2D PCA projection is used. Different classes are plotted using different colours and checked if the classes are well-separated from each other.

Figure 58 PCA Analysis



Source: own

```
array([0.02523481, 0.02446406])
```

The explained variance in Figure 58 shows how much information can be attributed to each of the components which is 0.02.

Linear dependence of Charged Off

On the training set, Pearson correlation is computed for each predictor with the response variable Charged Off. Sort the results by the absolute value of the Pearson Correlation.

Table 20 Pearson Correlation of the Charged Off loans by abs value

0	int_rate	0.005013	5	term	0.003120
1	earliest_cr_line	0.004391	6	revol_util	0.002944
2	initial_list_status_w	-0.004307	7	dti	-0.002922
3	purpose_wedding	0.003933	8	addr_state_IA	0.002744
4	mort_acc	0.003478	9	sub_grade_A3	0.002569

Source: own

The variables most linearly correlated shown in Table 20 with the Charged Off are the int_rate, earliest_cr_line, initial_list_status_w, purpose_wedding, mort_acc, term, revol_util, dti, addr_state_IA, sub_grade_A3. It seems like the borrower's state of residence, the revolving balance, and the loan purposes are irrelevant for predicting Charge Off.

5.2 Modelling

Machine Learning algorithms are implemented and consist of one or more of the following steps, depending on the particular model:

1. Mean imputation of missing values
2. Dimension reduction using linear discriminant analysis (LDA)
3. Data standardization: rescaling to zero mean and unit variance
4. The chosen model

Evaluation and comparison of the following models using a cross-validated AUROC score and confusion matrix on the training set:

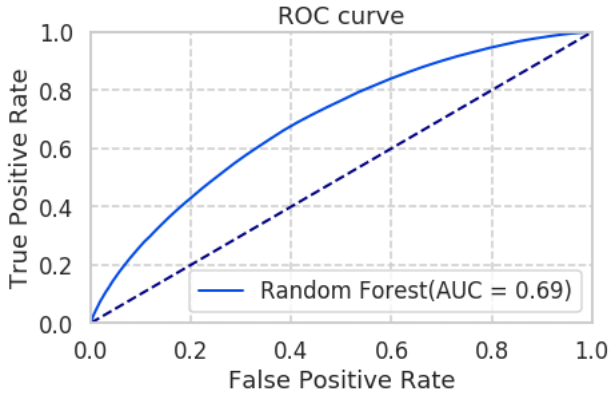
1. Logistic regression with SGD training
2. Random forest
3. K-nearest neighbors

5.2.1 **Random Forest**

Next random forest model is trained. Note that data standardization is not necessary for a random forest. The random forest takes very long to train, so different hyperparameter choices are most tested. GridSearchCV is used for consistency.

```
Best score:  0.7005000937215417
Best parameters set:
 {'class_weight': {0: 1, 1: 1}}
```

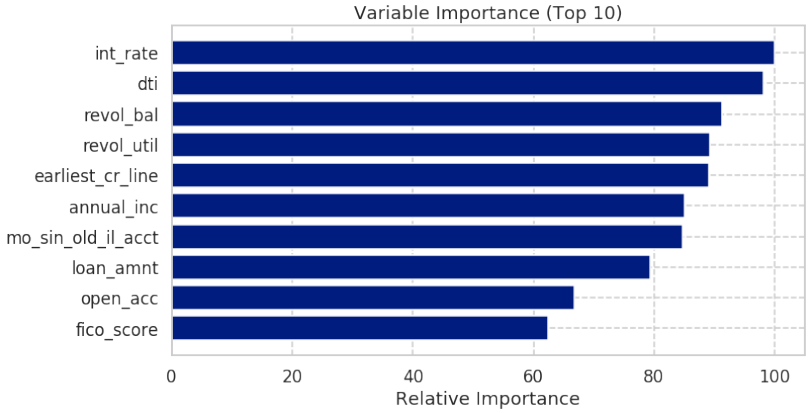
Figure 59 ROC Curve Random Forest training set



Source: own

The ROC curve for the Random Forest model is shown in the Figure 59 and its value is 0.69 which is considered to be quite decent.

Figure 60 Variable Importance Random Forest



Source: own

Looking at the variable importance, Figure 60, the most important ones is the int_rate, dti and revol_bal which have different weight comparing to the ones from the Pearson Correlation.

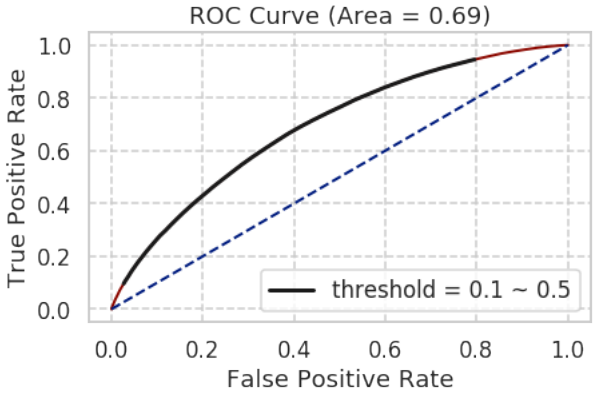
Visualizing a Single Decision Tree
See (Appendix 1).

Evaluation of the model on Training and Test Set

```
===== Random Forest =====
--- ROC AUC ---
Training Set: 1.0
Test Set: 0.6905067189345678

--- Accuracy ---
Training Set: 0.999986821966652
Test Set: 0.7825904091384625

--- Confusion Matrix ---
True Positive: 5325
False Negative: 56384
True Negative: 216098
False Positive: 5129

--- Precision ---
Training Set: 1.0
Test Set: 0.5093744021427205

--- Recall ---
Training Set: 0.9999323181049069
Test Set: 0.08629211298189891

--- F1 Score ---
Training Set: 0.999966157907205
Test Set: 0.1475825561575877
```

The values obtained in the Evaluation part are going to be analysed in detail in Conclusion part of this thesis.

Figure 61 ROC curve Random Forest on the test set



Source: own

The ROC curve for the Random Forest model on the Test set is shown in the Figure 61 and its value is 0.69 which is the same performance like the Training Set from Figure 59.

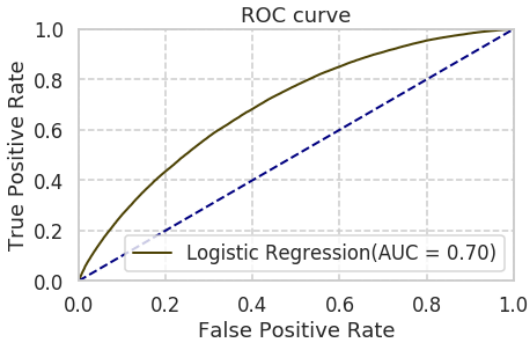
5.2.2 Logistic Regression with SGD training

Logistic regression with stochastic gradient descent training. The SGDClassifier estimator in scikit-learn implements linear classifiers (SVM, logistic regression, and others) with stochastic gradient descent (SGD) training. A particular linear classifier is chosen through the loss hyperparameter. Because the probability of Charged Off is predicted, logistic regression is chosen by setting loss = 'log'. Appendix C includes the mean cross-validated AUROC score of the best model.

The model performs as following:

```
Best score:  0.7117398669112628
Best parameters set:
 {'model__alpha': 0.1, 'model__penalty': 'l2'}
```

Figure 62 ROC Curve Logistic Regression



Source: own

The ROC curve for the Logistic Regression model is shown in the Figure 62 and its value is 0.7 which is slightly better than the Random Forest.

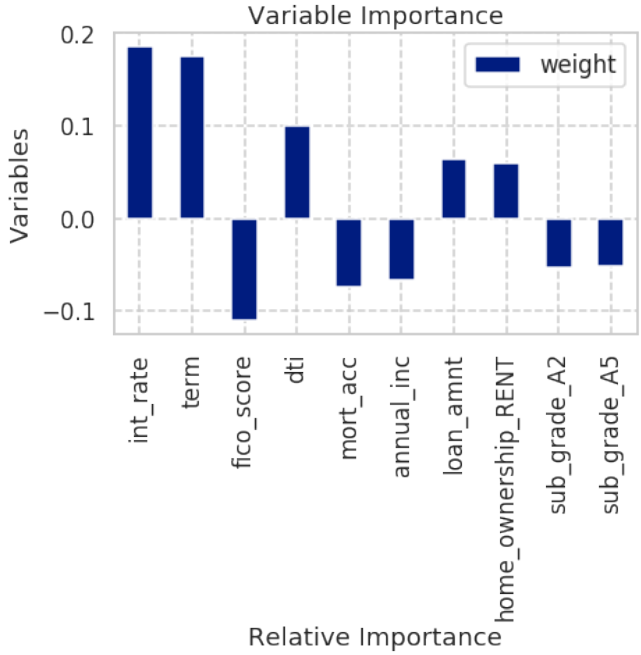
Table 21 Feature Importance

weight	
int_rate	0.186
term	0.175
fico_score	-0.111
dti	0.100
annual_inc	-0.074
mort_acc	-0.067
home_ownership_RENT	0.064
loan_amnt	0.059
sub_grade_A2	-0.053
sub_grade_A4	-0.052

Source: own

The values in Table 21 are slightly different than the ones from the previous Pearson Correlation analysis and Random Forest.

Figure 63 Variable Importance Logistic Regression



Source: own

Feature coefficients Figure 63 shows the top 5 features and their weights in the fitting are: Term (0.18), int_rate(0.156), dti(0.125), fico_score(-0.113), annual_inc(-0.093) while, another way to look at the features, RFE, shows the top 5 features are: int_rate, term, dti, fico_score, mort_acc. These two methods have slightly different ranks on annual_inc & mort_acct; int_rate and term.

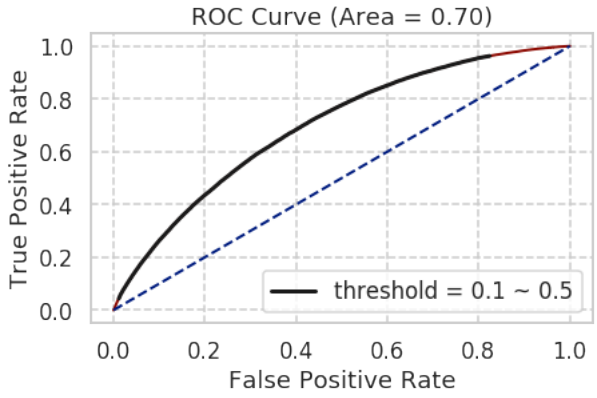
Evaluation of the model on Training and Test Set

```
===== Logistic Regression =====
--- ROC AUC ---
Training Set: 0.7138087454641363
Test Set: 0.6960010370631621

--- Accuracy ---
Training Set: 0.8066236560759206
Test Set: 0.7824525687788051

--- Confusion Matrix ---
True Positive: 2691
False Negative: 59018
True Negative: 218693
False Positive: 2534
```

```
--- Precision ---
Training Set: 0.6328066215199398
Test Set: 0.515023923444976

--- Recall ---
Training Set: 0.01626299250664733
Test Set: 0.04360790160268356

--- F1 Score ---
Training Set: 0.031711018730616596
Test Set: 0.08040756566169659
```

The values obtained in the Evaluation part are going to be analysed in detail in Conclusion part of this thesis.

Figure 64 ROC curve Logistic Regression test set



Source: own

The ROC curve for the Logistic Regression model on the Test set is shown in the Figure 64 and its value is 0.70 which is the same performance like the Training Set from Figure 62.

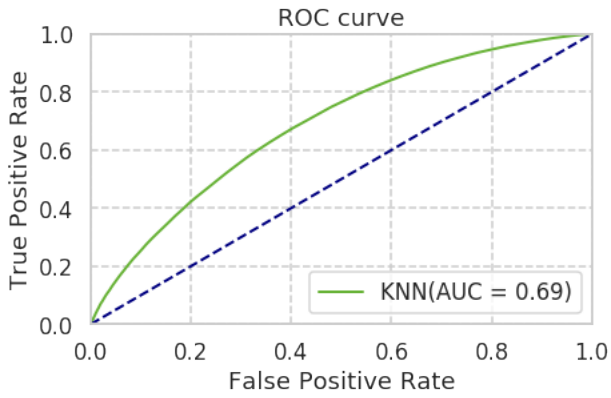
5.2.3 **KNN (k-nearest neighbors)**

Next is k-nearest neighbors. The number of variables is reduced to 10 for kNN to perform well. LDA for dimension reduction is used. The number of component variables to keep is a hyperparameter.

```
Best score:  0.7048576256360988
Best parameters set:
 {'lda__n_components': 3, 'model__n_neighbors': 125}
```

Only 3 LDA components are necessary for kNN to perform almost as well as logistic regression.

Figure 65 ROC Curve KNN



Source: own

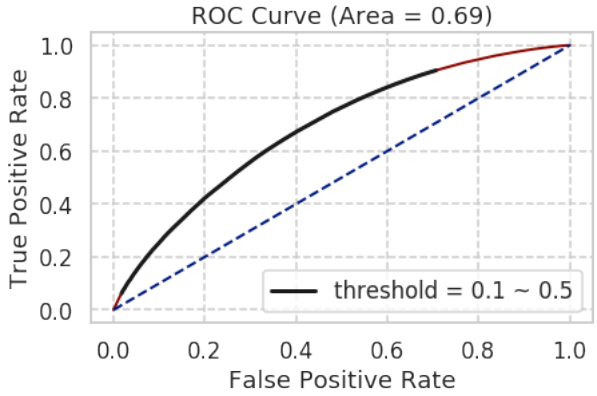
The ROC curve for the KNN model is shown in the Figure 65 and its value is 0.69 which is considered to be quite decent.

Evaluation of the model on Training and Test Set

```
===== KNN =====
--- ROC AUC ---
Training Set: 0.7226557852868281
Test Set: 0.6869660902345743

--- Accuracy ---
Training Set: 0.8082831469896665
Test Set: 0.7818411230808381

--- Confusion Matrix ---
True Positive: 3464
False Negative: 58245
True Negative: 217747
False Positive: 3480

--- Precision ---
Training Set: 0.5563003156364152
Test Set: 0.4988479262672811

--- Recall ---
Training Set: 0.07583272903069857
Test Set: 0.056134437440243726

--- F1 Score ---
Training Set: 0.13347117810480455
Test Set: 0.100913288567142
```

The values obtained in the Evaluation part are going to be analysed in detail in Conclusion part of this thesis.

Figure 66 ROC curve KNN test set

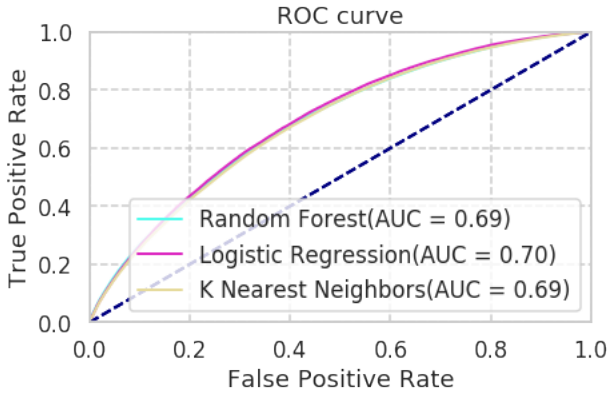


Source: own

The ROC curve for the KNN model on the Test set is shown in the Figure 65 and its value is 0.69 which is the same performance like the Training Set from Figure 66.

5.3 Comparative Results Overview

Figure 67 ROC Curve Comparative



Source: own

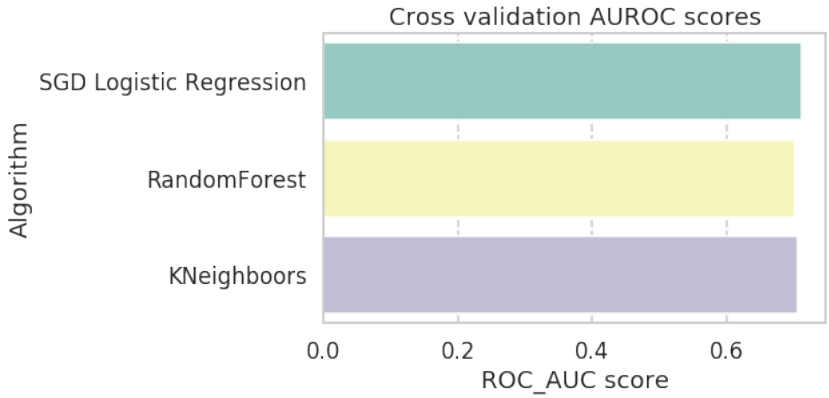
The ROC curve from Figure 67 shows that the Logistic regression squeaked out ahead and coupled with the fact that the SGDClassifier trains much faster than the other two models. Even if it performs only slightly better from the ROC curve perspective, it is the best out of these 3 models as it has other advantages such as processing time.

Table 22 AUROC Values

	AUROC	Algorithm
0	0.711740	SGD Logistic Regression
1	0.700500	RandomForest
2	0.704858	KNeighboors

Source: own

Figure 68 Cross validation AUROC scores



Source: own

Table 22 and Figure 68 show that from all three models the SGD Logistic Regression performs better. There were two functions written in the Evaluation part of each model which were run on both the train and test sample. One for summary of the performance of each model which includes the confusion matrix and the confusion matrix performance parameters and the other one for plotting the ROC for the test data.

As a conclusion, Logistic Regression performs slighter better than others, but it runs much faster. The Logistic Regression algorithm correctly predicts 221,384 instances out of 282,936 instances. Here the classification accuracy on the testing set is 78,24%. The algorithm correctly classifies 2691 instances as a Charged-off from a total number of 5225 which were actually Charged-off. On the other hand, it correctly classifies 218,693 instances as Fully Paid from a total number of 277,711. It has a classification accuracy of 4,3%.

6 Conclusion

Machine Learning techniques were used to predict the probability that a requested loan on Lending Club will be Charged Off.

After training with 3 models (Random Forest, Logistic Regression and KNN), all three models performed similarly according to a cross-validated AUROC score on the training data. Logistic Regression (with stochastic gradient descent) is chosen as the best model because it was the fastest model to train with the highest AUROC score of 0.71 and the score on the testing dataset is slightly lower at 0.70. This model, while far from perfect, can provide a somewhat informed prediction of the likelihood that a loan will Charge Off, using only data available to potential investors before the loan is fully funded.

Finally, from Figure 63 we can see that the most important feature is the interest rate. This is not a surprise because this feature reflects the behaviour of the borrower and it is classified by their credit history. The interest rate is determined by the assigned grade. Higher rates are related to more risky loans. The purpose of loan is also very important. According to the Pearson correlation between the features and the target variable, the most important features for prediction of the Charged-off loans are: interest rate, term, debt-to-income ratio, and FICO score.

According to the Evaluation section of each model the following conclusions can be made:

The Logistic Regression algorithm correctly predicts 221,384 instances out of 282,936 instances. Here the classification accuracy on the testing set is 78,24%. The Random Forest algorithm correctly predicts 221,423 instances out of 282,936 instances. Here the classification accuracy on the testing set is 78,25%. The KNN algorithm correctly predicts 221,211 instances out of 282,936 instances. Here the classification accuracy on the testing set is 78,18%.

As per Logistic Regression, the algorithm correctly classifies 2691 instances as a Charged-off from a total number of 5225 which were actually Charged-off. On the other hand, it correctly classifies 218,693 instances as Fully Paid from a total number of 277,711. It has a classification accuracy of 4,3%. As per KNN, the algorithm correctly classifies 3464 instances as a Charged-off from a total number of 6944 which were actually Charged-off. On the other hand, it correctly classifies 58245 instances as Fully Paid from a total number

of 275,992. It has a classification accuracy of 5,6%. As per Random Forest, the algorithm correctly classifies 5325 instances as a Charged-off from a total number of 10454 which were actually Charged-off. On the other hand, it correctly classifies 56384 instances as Fully Paid from a total number of 272,482. It has a classification accuracy of 8,6%.

Through experiments, the model which best suits the dataset and serves the purpose of giving an investor a model which would increase their chances of a profit was found. It had an accuracy of 0.78 and a precision of 0.51. The investor might pass on a lot of loan opportunities, but there are very less chances of losing money. Although the overall prediction accuracy is good, the prediction accuracy of Charged Off instances is not that good. The major reason for this could be the class imbalance i.e. high number of instances having class as 'Fully Paid', which results in biased output.

While the models that have a decent prediction accuracy and give larger returns on profit than the baseline model were found, the increase in performance is rather small. Since the univariate interest rate model performs comparably well, it indicates that the credit scoring done by LendingClub is fairly accurate. The interest rate thus constitutes a decent indicator of the risk involved with the loan, and the improvements from using the learning techniques are somewhat limited.

Despite the incredible growth in P2P Lending, many Americans still hesitate to invest on these platforms, perhaps due to fear, barriers to entry or perceived complexity. This thesis, above all, seeks to dispel the idea that Lending Club investment strategies are fundamentally distinct from the stock market or from investing in bonds. They are not. Like bonds or like stocks, they simply reflect an expected stream of payments which may or may not actualize, and our framework treats them as such. In conclusion, this thesis does not intend to encourage readers to invest their life savings in a 60-month filtered equal allocation or even, as many have done, to invest in online P2P Lending as an alternate investment strategy to benefit the entirety of one's portfolio. Rather, I simply wish to demonstrate a framework for analysis and a description of the feature set that could be used to better understand and appreciate this ground-breaking investment platform. The new age of personal credit has arrived, and we should be wise to at least take it seriously.

7 References

Cielen, D., Meysman, A. & Ali, M., 2016. *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools.* 1 ed. Shelter Island: Manning Publications.

Copeland, M., 2016. *What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?.* [Online]
Available at: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/
[Accessed 10 9 2019].

Cunningham, S., 2019. *Lending Club vs. Prosper 2019: 5 Big Differences.* [Online]
Available at: https://www.lendingmemo.com/lending-club-vs-prosper-for-borrowers/
[Accessed 10 8 2019].

Abbott, D., 2014. *Applied Predictive Analytics.* 1 ed. San Diego: Wiley.

Akopian, A., 2016. *Технологии Big Data в экономике. От теории к фактам..* [Online]
Available at: https://itcrumbs.ru/texnologii-big-data-v-ekonomike_18032
[Accessed 10 8 2019].

Anderson, R., 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation.* 1 ed. London: Oxford University Press.

Bari, A., 2016. *Predictive Analytics For Dummies.* 2 ed. New Jersey: For Dummies.

Breimann, L., 2001. Random Forests. *Machine Learning,* 45(1), pp. 5-32.

Dietterich, T. G., 1997. Ensemble Learning. *AI Magazine.*

Dietterich, T. G., 2000. Ensemble Methods in Machine Learning.. *Multiple Classifier Systems*, pp. 1-15.

Domingos, P., 2012. A Few Useful Things to Know About Machine Learning. *Communications of the acm,* 55(10), p. 78.

Gandhi, R., 2018. *K Nearest Neighbours — Introduction to Machine Learning Algorithms.* [Online]
Available at: https://towardsdatascience.com/k-nearest-neighbours-introduction-to-machine-learning-algorithms-18e7ce3d802a
[Accessed 25 9 2019].

Gandhi, R., 2018. *Support Vector Machine — Introduction to Machine Learning Algorithms.* [Online]
Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
[Accessed 24 9 2019].

George, N., 2019. *All Lending Club loan data.* [Online]
Available at: https://www.kaggle.com/wordsforthewise/lending-club
[Accessed 20 3 2019].

Hayes, A., 2019. *FICO Score.* [Online]
Available at: https://www.investopedia.com/terms/f/ficoscore.asp
[Accessed 10 8 2019].

He, H. & Garcia, E. A., 2009. Learning from Imbalanced Data. *Knowledge and Data Engineering*, 26 6, pp. 1263 - 1284.

Hurwitz, J. S., Nugent, A., Halper, F. & Kaufman, M., 2013. *Big Data For Dummies.* 1 ed. New Yersey: For Dummies.

Kagan, J., 2019. *Peer-To-Peer Lending (P2P).* [Online]
Available at: https://www.investopedia.com/terms/p/peer-to-peer-lending.asp
[Accessed 10 8 2019].

Kubat, M., 2015. *An Introduction to Machine Learning.* 2 ed. Cham: Springer.

Kuhn, M., 2018. *Applied Predictive Modeling.* 1 ed. New Yorl: Springer.

Lending Club, 2019. *Data Dictionary.* [Online]
Available at: https://help.lendingclub.com/hc/en-us/articles/216127307-Data-Dictionaries
[Accessed 10 10 2019].

Lending Club, 2019. *Understanding Credit Scores.* [Online]
Available at: https://www.lendingclub.com/loans/resource-center/understanding-credit-scores
[Accessed 10 8 2019].

Lending Club, 2019. *What do the different Note statuses mean?.* [Online]
Available at: https://help.lendingclub.com/hc/en-us/articles/215488038-What-do-the-different-Note-statuses-mean-
[Accessed 7 7 2019].

Lending Club, 2019. *What is the difference between a loan that is in "default" and a loan that has been "charged off"?.* [Online]
Available at: https://help.lendingclub.com/hc/en-us/articles/216127747
[Accessed 7 7 2019].

Loughlin, R., 2015. *LendingClub: Opening the Floodgates on Credit.* [Online]
Available at:
https://www.hbs.edu/openforum/openforum.hbs.org/goto/challenge/understand-digital-transformation-of-business/lending-club-opening-the-floodgates-on-credit/comments-section.html
[Accessed 10 8 2019].

Lustman, S., 2013. *Big Data & P2P Lending: Don't Be Misled.* [Online]
Available at: https://p2plendingexpert.com/big-data-p2p-lending-dont-be-misled/
[Accessed 10 8 2019].

Lustman, S., 2017. *P2P Investing 101: Why the Smart Money Invests In Peer to Peer Loans.* 1 ed. Maryland: CreateSpace Independent Publishing Platform.

Manyika, J. et al., 2014. *Global flows in a digital age: How trade, finance, people, and data connect the world economy,* New York: McKinsey Global Institute.

Marsland & Stephen, 2011. *Machine Learning: An Algorithmic Perspective.* 1 ed. Palmerston North: Chapman and Hall/CRC.

MathWorks, 2019. *Predictive Analytics 3 Things You Need to Know.* [Online]
Available at: https://www.mathworks.com/discovery/predictive-analytics.html
[Accessed 10 8 2019].

Mitchell, T. M., 2006. *The Discipline of Machine Learning.* [Online]
Available at: http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf
[Accessed 10 9 2019].

Navlani, A., 2018. *Decision Tree Classification in Python.* [Online]
Available at: https://www.datacamp.com/community/tutorials/decision-tree-classification-python
[Accessed 24 9 2019].

Nisbet, R., Elder IV, J. & Miner, G., 2009. *Handbook of Statistical Analysis and Data Mining Applications.* 1 ed. London: Academic Press.

O'Rourke, T., 2016. *Lending Club - Predicting Loan Outcomes.* [Online]
Available at: https://rpubs.com/torourke97/190551
[Accessed 30 9 2019].

Pyle, D. & San José, C., 2015. *An executive's guide to machine learning.* [Online]
Available at: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/an-executives-guide-to-machine-learning#
[Accessed 10 9 2019].

Rind, V., 2016. *Pros and Cons of Peer-to-Peer Lending.* [Online]
Available at: https://www.gobankingrates.com/loans/personal/5-perks-peer-to-peer-lending/
[Accessed 10 8 2019].

Semenov, A., 2019. *Big Data Technologies.* [Online]
Available at: https://www.uplab.ru/blog/big-data-technologies/
[Accessed 10 8 2019].

Sertic, B., 2017. *LOAN RISK PREDICTION: AN APPLICATION FOR INVESTORS AT LENDING CLUB.* [Online]

Available at: https://api.rpubs.com/boris_sertic/lrpm
[Accessed 9 9 2019].

Stanford, 2019. *Machine Learning.* [Online]
Available at: https://www.coursera.org/learn/machine-learning
[Accessed 10 7 2019].

Wikipedia, 2019. *Zopa.* [Online]
Available at: https://en.wikipedia.org/wiki/Zopa
[Accessed 10 8 2019].

Woodruff, M., 2014. *Here's what you need to know before taking out a peer-to-peer loan.* [Online]
Available at: https://finance.yahoo.com/news/what-is-peer-to-peer-lending-173019140.html?guccounter=1
[Accessed 10 8 2019].

Wu, J., 2014. *Loan default prediction using lending club data.* [Online]
Available at: http://www.wujiayu.me/assets/projects/loan-default-prediction-Jiayu-Wu
[Accessed 24 9 2019].

8 Appendix

8.1 Appendix A

Table 23 Variable Description

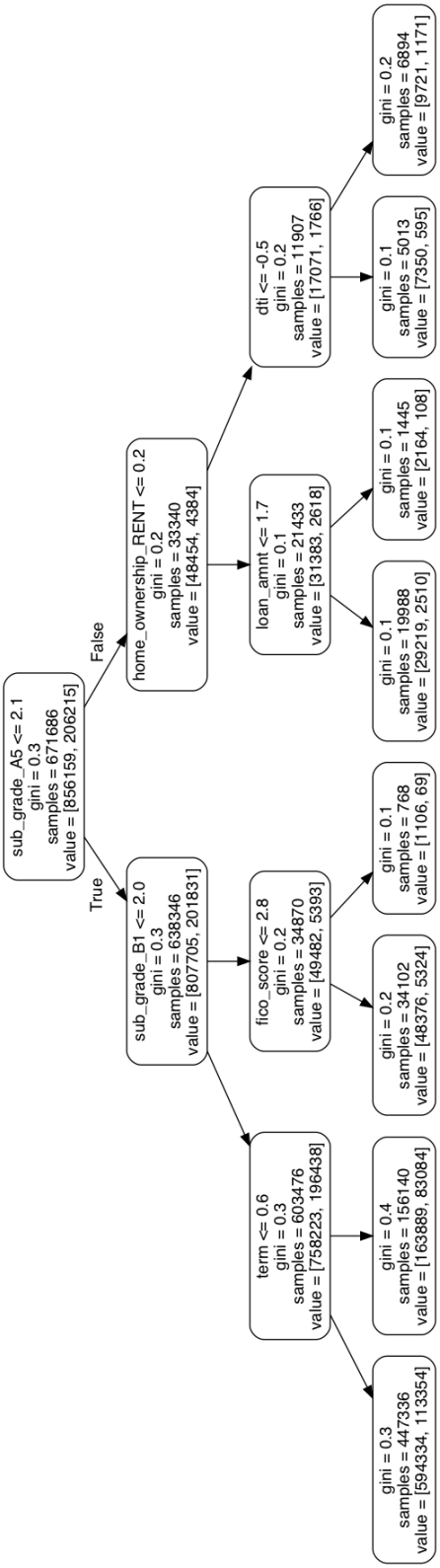
Transformed	Feature	Description
Yes	addr_state	The state provided by the borrower in the loan application
No	annual_inc	The self-reported annual income provided by the borrower during registration.
Yes	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
No	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requeste LC loan, divided by the borrower's self-reported monthly income.
No	earliest_cr_line	The month the borrower's earliest reported credit line was opened
Yes	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Yes	emp_title	The job title supplied by the Borrower when applying for the loan.*
Yes	fico_range_high	The upper boundary range the borrower's FICO loan origination belongs to.
Yes	fico_range_low	The lower boundary range the borrower's FICO loan origination belongs to.
Yes	grade	LC assigned loan grade
Yes	home_ownership	The home ownership status provided by the borrower during registration or obtained from th credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
No	id	A unique LC assigned ID for the loan listing.
Yes	initial_list_status	The initial listing status of the loan. Possible values are – W, F
No	installment	The monthly payment owed by the borrower if t loan originates.
	int_rate	Interest Rate on the loan
No	issue_d	The month which the loan was funded
No	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
Yes	loan_status	Current status of the loan
No	mo_sin_old_il_acct	Months since oldest bank installment account opened
No	mo_sin_old_rev_tl_op	Months since oldest revolving account opened
No	mort_acc	Number of mortgage accounts.
No	open_acc	The number of open credit lines in the borrower' credit file.

No	policy_code	publicly available policy_code=1 new products not publicly available policy_code=2
No	pub_rec	Number of derogatory public records
No	pub_rec_bankruptcies	Number of public record bankruptcies
Yes	purpose	A category provided by the borrower for the loa request.
No	revol_bal	Total credit revolving balance
No	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
Yes	sub_grade	LC assigned loan subgrade
No	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
No	title	The loan title provided by the borrower
No	total_acc	The total number of credit lines currently in the borrower's credit file
No, excluded	url	URL for the LC page with listing data.
Yes	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
Yes	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

Source: (Lending Club, 2019)

8.2 **Appendix B**

Figure 69 Small Decision Tree - Random Forest



Source: own

8.3 **Appendix C**

Table 24 Mean cross-validated AUROC score of the best model
Source: own

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_model__alpha	param_model__p	params	split0_test_score	split1_test_score	split2_test_score	mean_test_score	std_test_score	rank_test_score
0	61.169369	2.342201	1.248937	0.577687	1e-05	l1	{'model__alpha': 1e-05, 'model__penalty': 'l1'}	0.704502	0.699550	0.677402	0.693818	0.011782	4
1	35.220870	11.072968	1.398240	1.111600	1e-05	l2	{'model__alpha': 1e-05, 'model__penalty': 'l2'}	0.706864	0.699004	0.676181	0.694017	0.013013	3
2	12.490705	2.904569	0.325992	0.058549	0.1	l1	{'model__alpha': 0.1, 'model__penalty': 'l1'}	0.500000	0.500000	0.500000	0.500000	0.000000	5
3	6.843454	0.297149	0.423599	0.169512	0.1	l2	{'model__alpha': 0.1, 'model__penalty': 'l2'}	0.732133	0.719773	0.690160	0.714022	0.017611	1
4	11.103602	1.085517	0.897483	0.185883	100	l1	{'model__alpha': 100, 'model__penalty': 'l1'}	0.500000	0.500000	0.500000	0.500000	0.000000	5
5	9.761820	2.193582	0.592456	0.305512	100	l2	{'model__alpha': 100, 'model__penalty': 'l2'}	0.726649	0.714085	0.668761	0.703165	0.024862	2