Credit Risk Analysis with Machine Learning for Peer-to-Peer Lending

Lingzhi Huang (41710) & Wanlin Hu (41692)

Abstract

In the past decade, the scale and scope of fintech credit have snowballed. The peer-to-peer lending industry can be seen as a complement to the traditional banking system. Improving the performance of lending platforms by increasing the accuracy of credit default predictions can help these platforms establish a decisive advantage in the market. This thesis aims to investigate the application of machine learning techniques to P2P lending default prediction modelling. It will seek to identify the most optimal approach for default prediction using machine learning for a given evaluation metric. This study uses real loan data from LendingClub, a publicly accessible public data source, to conduct its credit analysis. A well-rounded set of evaluation metrics was carefully designed and compared. This study discusses four well-established machine learning techniques: logistic regression, support vector machine, random forest, and K-nearest neighbour algorithm. Logistic regression is considered the most adaptable approach for P2P default estimation among the available evaluation metrics after analysing the modelling results. This thesis is of great relevance in helping P2P platforms to improve their ability to identify the credit risk of borrowers. It can also help improve the success rate of P2P online lending, promote reasonable and effective investment by lenders, and boost the development of the P2P online lending industry.

Keywords: Credit Risk, Default, Machine Learning, Peer-to-Peer Lending, Logistic Regression, SVM, Random Forest, k-NN, Predictive Modelling

Stockholm School of Economics

Department of Finance

Master Thesis in Finance

Spring 2021

Supervisor: Tobias Sichert

Acknowledgements

We would like to sincerely appreciate our supervisor Assistant Professor Tobias Sichert for his support and instruction during the writing process. His knowledge and insights genuinely provided us with a more profound and logical understanding of our research topic.

Stockholm, May 2021

Lingzhi Huang & Wanlin Hu

List of Figures

List of Tables

Table of Contents

Chapter 1 Introduction

The aim of this thesis is presented. This chapter includes the study background, the research question, and details the structure of the thesis.

1.1 Background

In the past decade, the scale and scope of fintech credit have increased dramatically. The first batch of online loan platforms targeted the unsecured consumer credit market, focusing on borrowers with insufficient bank services. Since then, online lending platforms have developed and expanded to other markets, including student loans, auto loans, mortgage loans, and SME corporate loans. According to Tang's (2018) research on the U.S. financial technology credit market, online loans provide an alternative to uncollateralised lenders. Financial technology credit services can also supplement bank loans in catering to the demand of underserved borrowers.

Zopa, the first P2P lending platform, was founded in the UK in 2005. From there, the concept spread globally. According to Bertsch (2019), the UK and the US are the largest markets in the developed world, while China is the most active market for P2P credit. A survey by the Cambridge University Centre for Alternative Finance (2018)suggests that over 75% of global fintech credit activity occurs in China, with over a thousand active online lending platforms in the country. In the US, Prosper launched in 2006 followed by the LendingClub in 2007, and these companies remain the country's two largest P2P lending platforms. In 2008, the U.S. Securities and Exchange Commission (SEC) officially confirmed that P2P transactions in the form of securities are regulated under the Securities Act of 1993. From 2012 to 2015 the P2P industry expanded rapidly, with the number of platforms peaking at 13,000 in December 2014 according to the CEIC database (2021). After the 2008 financial crisis, P2P lending was advantageous for both lenders and borrowers, attracting investors who were frustrated with stock market returns and the low interest rates offered by banks (Brennan, 2009). These platforms also attracted large institutional investors, such as hedge funds and wealth management firms (Light, 2012). The 'Global Peer-to-Peer Lending Industry' report

(Reportlinker.com, 2020) predicted that the global peer-to-peer lending market will grow at a compound annual growth rate of 42.7% between 2020 and 2027.

One of the main functions of finance is to optimise the allocation of social resources. Traditional bank credit assessment models rely heavily on financial records, collateral assets or government guarantees. However, SMEs, entrepreneurs and individuals in urgent need of capital often cannot fulfil these conditions, leaving investors with idle funds. Peer-to-peer (P2P) lending bridges the gap created by traditional financial models by allowing private individuals to offer small loans directly to private borrowers or SMEs, generally without collateral. This improves the utilisation rate of idle capital and solves the problem of difficult and long financing cycles for those in need. It also enables small businesses and entrepreneurs, who have difficulty obtaining credit from traditional financial institutions such as banks, to have a wider source of funding, supporting SMEs and increasing employment.

Although P2P online lending has advantages that traditional lending models do not have, there are also many risks. In general, these into two categories: platform risk and borrower credit risk. One of the key risks to the platform is the low barrier to entry into the P2P network lending industry due to initial confusion in the market and a lack of regulatory policies. Many platforms have incorrect market positioning, unreasonable risk control and irregularities in their accounting for lending transactions, leading to losses for both the platform and the lender. There are even cases of platforms engaging in fraud, illegally holding on to the money of their customers.

However, the credit risk of borrowers is the most significant risk faced by P2P network lending. This can also be described as default risk, reflecting the issue of a borrower who is unable to fulfil the obligations of the lending agreement, resulting in losses to stakeholders such as the platform and the lender. Online lending default prediction is a technique for managing the risks of online lending. The large amount of real-world online lending data in the peer-to-peer industry offers the possibility of implementing default prediction. Therefore, finding a method with optimal performance in accurately assessing and predicting the default risk of borrowers is crucial to the stable operation of P2P online lending, the focus of this paper.

1.2 Research question

The purpose of this thesis is to explore the potential of applying machine learning techniques to P2P lending for default prediction modelling. This thesis is of great relevance given its potential to help P2P platforms improve their ability to identify the credit risk of borrowers. In turn, this could improve the success rate of P2P online lending, promote reasonable and effective investment by lenders, and accelerate the development of the P2P online lending industry. The research question is the following:

- *Among the selected machine learning methods, which method performs best in default prediction in peer-to-peer lending for a given model evaluation metric?*

1.3 Structure of the thesis

Credit risk is the main risk faced by P2P online lending platforms. Using the public data of LendingClub as an example, this thesis selects suitable variable indicators that affect borrowers' credit default risk. It then uses multiple machine learning models to conduct prediction research on borrowers' credit risk before conducting a comparative analysis of the prediction effects of different models. This analysis is based on the relevant indicators of the evaluation models, aiming to improve platform loan default prediction ability and reduce bad debts, lessons which can then be applied to future risk identification and analysis.

The structure of the thesis is as follows:

Chapter 1 presents the introduction, including a background to the P2P industry, highlights the research question, and lists the research methodology.

Chapter 2 is a literature review of P2P online lending. It assesses the key factors influencing P2P financing and the credit risk of borrowers through various risk assessment models.

Chapter 3 introduces the theoretical basis for the four models.The theories of Logistic Regression, Random Forest, Support Vector Machines (SVMs), and k-Nearest Neighbours (k-NN) are discussed.

Chapter 4 provides a comparative analysis of credit risk in P2P lending based on empirical data. The variable indicators in the dataset are screened and then different models are applied

to predict default. The prediction performance of each model is analysed and compared through the evaluation models. From there, the model with the best prediction effect is identified. R x64 3.6.3 and R version 3.6.1 on x86_64-apple-darwin15.6.0 (64-bit) are used to programme and analyse the results.

Chapter 5 offers a summary of the research problem. The results of the empirical and comparative analyses are brought together to form a set of preliminary conclusions. The imperfections of the article's analysis and possible avenues for future research directions are then presented.

Chapter 2 Review of P2P Lending Research

As an emerging lending model, P2P lending has attracted the attention of many scholars, leading to a considerable body of published literature on the subject. In this chapter, a review of this research is presented.

The information asymmetry that results from the virtual nature of online P2P lending has led to the credit risk of borrowers becoming the most significant risk facing this market. As such, this issue has been extensively studied by scholars. Previous researchers have shown particular interest in two aspects of P2P lending: the determinants of P2P financing and the quantitative models used to predict credit default.

2.1 Determinants of P2P lending

Regarding the causes of credit risk in online finance, Yum et al. (2012) point out that information asymmetry is the most fundamental problem faced by the online lending model. On the one hand, investors cannot gain full knowledge of the credit risk of borrowers because there is no effective communication channel between investors and borrowers. On the other hand, borrowers on P2P platforms are often at a disproportionately high risk of default since they have turned to P2P lending because they are unable to obtain loans from traditional banks. This information mismatch also raises the issue of adverse selection. Investors want more comprehensive and trustworthy information about the borrower. However, borrowers have an incentive to withhold unfavourable information in order to qualify for a more favourable lending agreement, such as one with a lower interest rate. Platforms typically require borrowers to provide verifiable information such as financial records and demographic data, including gender, race and age, alongside non-verifiable information such as interests, family background and photographs. These characteristics are known as the determinants of P2P lending as they have a significant impact on whether a borrower receives the financing requested and the interest rate at which this is offered.

The majority of peer-to-peer lending sites provide lenders with a summary of the borrower's financial characteristics, which serves as a critical measure of their creditworthiness. Credit rating, income level, monthly expenditure, home ownership, and debt-to-income ratio are all

examples of typical financial characteristics. These are generally calculated by external rating agencies that compile a credit score from a variety of personal and financial characteristics. These financial determinants have been demonstrated to be crucial to the success of peer-to-peer financing (Iyer et al., 2009; Klafft, 2008; Freedman & Jin, 2008).

Research on whether there is age discrimination in the P2P lending market presents a consistent finding that younger and older people have more difficulty obtaining loans (Gonzalez & Komarova, 2014; Pope & Sydnor, 2008; Ravina et al., 2012). Analysing Prosper data, Ravina et al. (2012) found that older adults were discriminated against in the P2P loan market. Although older adults did not have higher default rates, on average they paid interest rates that were 14 basis points higher than the norm. Similarly, young people are considered a high-risk, high-default group. The 35–60-year-old group is 40-90 basis points more likely to obtain a loan compared to people under 35 (Pope & Sydnor, 2008).

There is no consistent finding for whether gender determines the success of P2P financing. Pope & Sydnor (2008) found male gender discrimination in the U.S. P2P online lending market. Although the expected return on investment for male borrowers was about 2% higher than that of single female borrowers, single female borrowers were more likely to be trusted by investors and to receive loans at a 0.4% lower interest rate. Similarly, Chen et al. (2017) found significant gender discrimination in China's P2P online lending market. After analysing transaction data collected on Chinese P2P online lending platforms from August 2007 to August 2011, they found that although women were more likely to obtain loans than men, women were less likely to use P2P online lending with only 20.64% of applicants being female. In this case, women had lower default rates than men but paid higher interest rates. In contrast, an analysis of transaction records from the German P2P lending platform Smava from March 2007 to March 2010 found that gender had little effect on borrowers' financing success after controlling for factors such as interest rate, the amount borrowed, loan term, loan purpose, financial status and age, and region of employment (Barasinska & Schäfer, 2014). Overall, there is no consensus on how gender affects the success of P2P financing, with empirical studies suggesting that the impact of gender varies between countries.

The completion of the loan description can introduce more personal information about the borrower and the goals of the loan. This can help to improve the borrower's credibility and convince lenders to offer a loan. The work of Herzenstein et al. (2008) suggests that

demographic attributes, such as race, gender, and marital status, have little impact on P2P fundraising success when compared to financial strength and advertising to potential lenders. In contrast, a study conducted by Larrimore et al. (2011) found that the words chosen in the loan description produced a significant effect. After studying more than 220,000 Prosper transactions, they found that loan descriptions with more words, specific descriptions (such as articles, quantifiers, and prepositions), number words, and words describing the borrower's ability to repay the loan all increased success rates. Conversely, loan descriptions that provided more human and contextual details, such as a description of friends, religion, and family, as well as more explanatory words (should/can/will) decreased the success rate for acquiring a loan.

The role of appearance in the P2P lending market is a new area of research and there are no firm conclusions yet. Some studies (Klafft, 2008; Pope & Sydnor, 2008) point out that lists with and without photos have almost the same interest rates but an increased likelihood of loan financing. However, other studies have concluded that good-looking borrowers are more likely to obtain loans and borrow at low interest rates (Ravina, 2007; Ravina et al., 2012).

Studies of borrower identification consistently conclude that more information is useful in determining a borrower's true creditworthiness, whether it is verified or unverified. Investors often infer the creditworthiness of borrowers based on facts, though it should be clear that the information used may often be non-standard and subjective (Iyer et al., 2009) This soft information is also widely used in decision-making for lower-level credit approvals. Given this, non-verifiable information disclosure can affect financing decisions just as much as objective, verifiable information (Yum et al., 2012).

2.2 Quantitative assessment of credit risk

Another common research theme explores the credit risk of borrowers through various credit risk assessment models. Scholars have been attempting to estimate credit risk for nearly a century. Following the early work of Fitzpatrick (1932), scholars began to use quantitative methods to assess the creditworthiness of consumers. To date, many methods have been employed to assess the creditworthiness of applicants, using either a traditional statistical approach or advanced machine learning techniques. Recent research has uncovered substantial evidence that machine learning methods can significantly improve the accuracy of statistical methods without relying on restrictive assumptions. Machine learning methods

refer to *'a set of algorithms specifically designed to tackle computationally intensive pattern-recognition problems in extremely large datasets*' (Khandani et al., 2010). Popular machine learning methods used in credit risk modelling include logistic regression (Steenackers & Goovaerts, 1989), artificial neural networks (Byanjankar et al., 2015; West, 2000), support vector machines (Huang et al., 2007; Van Gestel et al., 2003), decision trees (Zekic-Susac et al., 2004), discriminant analysis (Eisenbeis, 1978), and k-nearest neighbour analysis(Henley & j. Hand, 1997). These techniques are well suited for consumer credit risk analysis due to the large sample size and the complexity of the possible relationships between consumer's transactions and their characteristics (Khandani et al., 2010).

Myer and Forgy (1963) evaluated the credit risk of borrowers from retail credit application data to develop a scoring system for predicting credit risk. They conducted discriminant and multiple regression analysis, finding that at lower score levels, basic discriminant analysis outperforms multiple regression in separating customer groups and minimising potential default losses at a minimal cost. Wiginton (1980) compared linear and logit models in scoring experiments and found that the logit model outperforms the linear discriminant model. Most of the early modelling was based on simple algorithms, such as logistic regression and linear discriminant analysis. Later, with the development of data mining techniques, more data mining algorithms were used in risk control models for credit default prediction.

Moving past early machine learning methods, the application of advanced algorithms such as neural networks and support vector machines (SVM) has further improved the predictive ability of the models. Several studies have concluded that artificial neural networks (ANNs) are more accurate in their use of datasets than data analytics and linear regression (Abdou et al., 2008; Desai et al., 1996; Imtiaz & Brimicombe, 2017). However, many parameters such as network topology, learning rate, and training method must be fine-tuned before an ANN can be successfully deployed. SVM has since emerged as a competing method to ANN, with both a type of nonparametric method. SVM was proposed by Cortes and Vapnik (1995) as an object classification method that does not consider multicollinearity among predictors. The method uses kernel functions to transform input data into higher feature dimensions. As reported by Dong et al. (2015), this means that SVMs using the RBF kernel can perfectly classify datasets irrespective of data comparability. If the data comparability is low, the accuracy of the classification is proportionally low and fluctuates irregularly. In these cases, the data comparability is more important and useful to improve than adjusting the algorithm

or measurement parameters. Building on the explanation above, the purpose of this work is to understand the accuracy of SVM as a method for credit scoring. In recent years, ensemble algorithms such as Random Forest and Lightgbm have seen increasing use in default modelling. These use multiple learning algorithms to achieve better predictive performance than any individual algorithm can provide. The predictive power of the ensemble model is greater than models using simpler principles such as logistic regression and linear discrimination.
In addition, scholars have compared the performance of different machine learning methods to find more optimal methods for distinguishing loans with different risks.

However, there is no consensus on which machine learning models are the best for all data. Galindo and Tamayo (2000) compare the classification and regression tree (CART) model with other machine learning models such as neural networks and K-nearest neighbour classifiers for their ability to improve the identification of good borrowers. The results of the comparison show that the CART decision tree model outperforms other methods in predicting the risk profile of borrowers. Malekipirbazari and Aksakalli (2015) construct a random forest based classification method to assess the credit risk of borrowers. They argue that it outperforms FICO credit scores and the credit segmentation of LendingClub platforms in identifying low-risk borrowers. As part of this, they note that finding high FICO Borrowers with high scores and high LendingClub ratings does not necessarily result in low default rates. Jin and Zhu (2015) set their work apart from other scholars who categorise behaviour into defaulted and non-defaulted. In their analysis, they classified the lending behaviour on LendingClub into defaulted, near-defaulted, and paid-up (non-defaulted), using this framework to compare five models: two decision tree models, two neural network models, and one SVM model. Empirical studies show that the predictive performance of SVM, CART, and Multi-layer Perceptron are almost identical.

2.3 Comment on the current literature

In this paper, we have summarised scholars' work up to now on P2P online lending. Most of the studies discussed assess the determinants of P2P financing or the credit risk of borrowers through various credit risk assessment models. In addition, there are a few comparative analyses of the ability of different machine learning models to predict borrowers' credit risk in P2P online lending. Many models have been studied but there is no definitive conclusion on which model performs best in predicting P2P loan defaults. There are multiple possible

explanations for this issue. First, different data sets and different data processing methods may deliver different results. Second, scholars often use different measurements as a benchmark to compare the performance of models, leading to differing conclusions.

Based on this summary of scholars' research, this paper applies the Logistic Regression, Random Forest, Support vector machines (SVMs), and k-Nearest Neighbours approaches to identifying borrowers' credit risk in P2P online lending, using data from the US online lending platform LendingClub. Studying the credit risk of borrowers in P2P network lending involves making judgments of whether borrowers can perform on time. This is essentially a dichotomous problem. All four models can be used to analyse the problem of identifying risks, from which the relevant evaluation indicators of the models are then compared and contrasted.

Chapter 3 Methodology

This chapter describes four selected machine learning models that will be used to evaluate credit default. The models that we choose are Logistic Regression, Random Forest, Support vector machine (SVM), and k-Nearest Neighbours (k-NN). In the section below, we explain the theories behind these models.

Logistic Regression

Logistic regression is a machine learning method for binary classification prediction problems, based on linear regression. Linear regression describes the relationship between the predicting variable and the predictive variable in a linear manner:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)} \quad (1)$$

For classification problems, we would like to have a probability between 0 and 1. Therefore, we introduce a sigmoid mapping of the output (Figure 3.1). This function can map any value between 0 and 1.

$$logistic(\tau) = \frac{1}{1 + \exp(-\tau)} \quad (2)$$

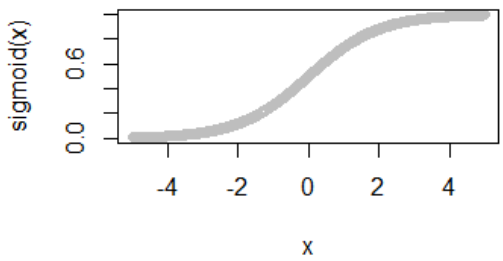

Figure 3.1: Plot of the sigmoid function

The linear regression expression assumes a sigmoid mapping, producing the following expression:

$$P\left(y^{(i)} = 1\right) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)} \quad (3)$$

As logistic regression will produce an output of between 0 and 1, this can be considered as a probability value. Therefore, a probability threshold can be taken, and a partition function constructed from the ratio of the output value to the threshold for achieving binary classification. For example, in P2P default prediction, the output of the model is the probability of a user defaulting in the future. If the threshold is 0.5, the user will be considered to have defaulted with a result greater than or equal to 0.5, with a dependent variable of 1. When the probability value is less than 0.5, the dependent variable is 0. This threshold can be set freely according to the situation.

Logistic regression uses a probabilistic framework, the maximum likelihood estimation, to form the parameters. Maximum likelihood estimation can only be discussed in an asymptotic context, meaning the model will only work well when the sample size is large enough. Conversely, when the sample size is relatively small, or the data distribution is unbalanced by many explanatory variables, the premises of the maximum likelihood method cannot be met, likely leading to biased estimates. The essence of logistic regression is that the probability distribution of the target variable must be assumed, after which a likelihood function is defined. The model construction of logistic regression is achieved by seeking the best parameter solution to maximise the value of this function through the iterative gradient method.

Logistic regression is simple, the model is highly interpretable, and by weighting the features you can see the effect of different factors on the result. However, logistic regression is sensitive to multicollinearity and also exhibits difficulty dealing with a data imbalance. Moreover, logistic regression itself cannot filter features and is not very accurate, due to its simple form, making it a poor fit for the real distribution of the data.

Random Forest

Random forest is a supervised learning algorithm in which decision trees are ensembled for classification, regression, and other tasks. Each decision tree in the forest is independent of the others. A decision tree is a structure in which each internal node represents a judgement on an attribute, each branch represents the output of a judgement, and each leaf node represents a classification result.

The random forest algorithm has a two-step process (Figure 3.2). First, it selects data points from the training set and uses them to build decision trees. Once the forest is formed, new data are input, and each decision tree individually determines which category the sample should belong to. In our case, it is whether the specific loan will default. Then the algorithm predicts which category the sample should belong to by observing which category has been selected the most.

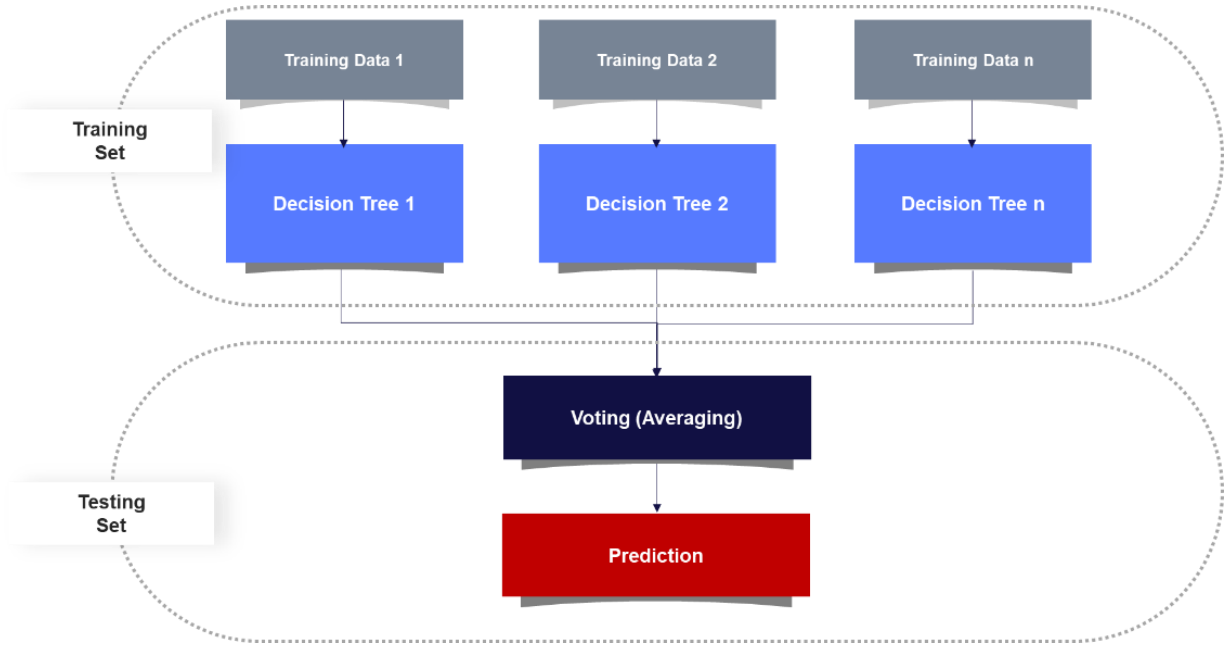

Figure 3.2 The two-step process for random forest

Random forest inherits the advantages of tree-based methods, and it is relatively easy to understand and explain. It is good at resolving errors in class-imbalanced data sets, and it also supports efficient methods to estimate missing data. This helps to maintain accuracy against overfitting even when there are large amounts of missing data. The disadvantage, however, is that features that take more divided values tend to have a greater impact on the decision-making process, affecting the validity of the fitted model. Compared with the simple tree method, the random forest is more difficult to interpret. This is its 'price' for higher predicting accuracy. In addition, although random forest can solve both classification and regression problems, it performs better in classification problems.

Support Vector Machine (SVM)

Support vector machines (SVMs) are a general class of linear classifiers that perform binary classification on data. They use supervised learning, with the decision boundary defined as the maximum marginal hyperplane over the learning sample. The maximum marginal hyperplane is the largest distance to the nearest element of each tag.

Figure 3.3 Best hyperplane for SVM

SVMs can be used for regression problems, but they are more used for classification. Each sample in the training set is identified as belonging to one of binary classes. The SVM constructs a model that categorises new samples, making it a non-probabilistic binary linear classifier. The SVM model represents samples as points in space, and that mapping allows samples from different classes to be separated by as wide a surface interval as possible. New samples are then mapped to the same space, and the class they belong to is predicted based on which side of the interval they fall. SVM can also apply to non-linear problems, and it is can cope with the linear non-differentiation of sample data. It does this mainly through relaxation variables and kernel function techniques.

The advantages of SVM are that it is effective in solving classification problems with high-dimensional features, that it still works well when the feature dimension exceeds the sample size, and that it is not dependent on the complete data set. It is also memory efficient, since only one subset of the training points is used in the decision process. SVMs are also versatile, leading to better performance in classification problems because class separation is often highly non-linear. The disadvantage is that SVMs perform poorly

when the feature dimension is much greater than the sample size, and they are less suitable for use when the sample size is very large and the kernel function is mapped to an extremely high dimension that is computationally demanding. Besides, the SVM is vulnerable to missing data.

K-Nearest Neighbors (k-NN)

The algorithm k-nearest neighbors (k-NN) is a non-parametric classification technique. Its fundamental concept is to add test data to known data and labels from the training set, compare the features of the test data to the corresponding features in the training set, and then find the top K data in the training set that are the most like it. The class corresponding to those test data is the one that occurs most frequently in those top K data.

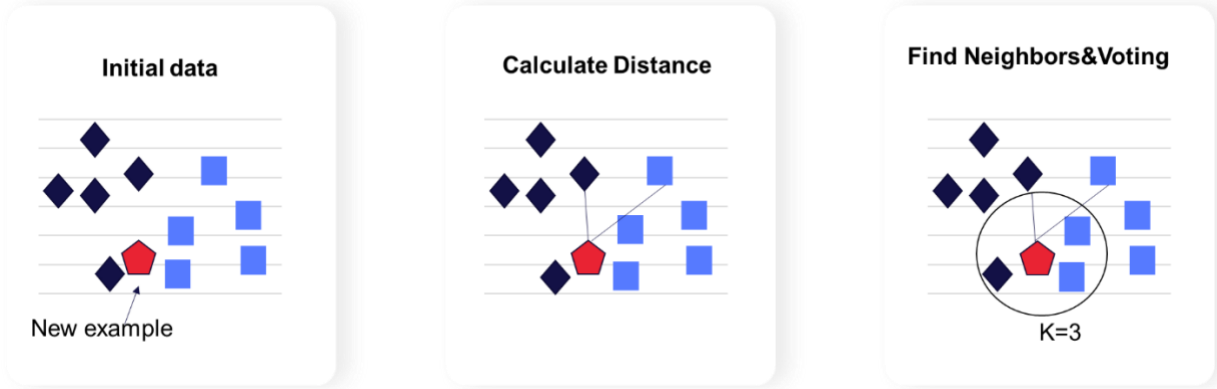

Figure 3.4 The k-NN approach

There are four major processes in the k-NN algorithm. First, it measures the distance between a sample point in the training and test samples. Then it ascends all distance values to pick the k samples with the smallest distance. The final move is to vote on the labels assigned to those k samples to determine the final classification category. The data are used to determine the optimal value of k. In general, a larger value of k in classification reduces the effect of noise, but it blurs the boundaries between categories. A better value of k can often be achieved using different heuristic techniques, such as cross-validation.

The advantage of k-NN is that it is intuitive and easy to understand and implement. It makes no assumptions about the data, so it is accurate and insensitive to outliers. Being a memory-based approach, k-NN immediately adapts to changes as new training data come in. This makes k-NN respond quickly to changes in inputs. However, unbalanced samples

can create more bias in predictions. In addition, for large datasets, the computation is intensive, so a large amount of memory is required.

Chapter 4 Empirical Analysis of Credit Risk in P2P Lending Market

Machine learning algorithms are used to conduct an empirical analysis of peer-to-peer credit data. The aim is to validate the role of machine learning algorithms in P2P credit default management. This section outlines the entire data mining process, from pre-processing the data and engineering features to creating the final model and evaluating the results. It uses four machine learning models as examples.

4.1 Source of Data – LendingClub

One characteristic of peer-to-peer lending is that P2P data are accessible for research purposes. Banking systems in the traditional sense have several layers of protection, and they are averse to disclosing private information. As a result, P2P lending would have access to a vast volume of data for credit risk assessment.

The data for the empirical analysis in this paper come from real loan application data on the LendingClub platform. LendingClub was founded in Delaware on October 2, 2006, and it was once the largest online lending marketplace website. It connected borrowers and investors in the United States. As of September 30, 2020, LendingClub's total loan originations exceeded $60 billion.

Under the promissory note model, LendingClub acts as the intermediary transferor of the loan. The borrower issues a promissory note for the loan to LendingClub, which originates the loan to the borrower and then assigns the promissory note issued by the borrower to the investor. Under the bank model, LendingClub works with WebBank, a commercial bank, which originates the loan to the borrower and assigns the loan promissory note issued by the borrower to LendingClub without recourse. LendingClub works with WebBank to address the issue of varying interest rate caps on loans across states and to eliminate the need to apply for state loan licenses. LendingClub assigns the loan promissory notes it receives from the borrower to the subscribing investor, who becomes a creditor of the borrower. Under the securities model, WebBank sells the loan to LendingClub after issuing it to the borrower, and then LendingClub sells the debt instrument to the investor. In this model, there is no direct debt relationship between the investor and the borrower. From these three changes, LendingClub's changes have

always been accompanied by changes in regulatory policies in the U.S. As an emerging fintech industry, the P2P industry is subject to strong and unstable policy constraints, which leads to higher compliance costs borne by P2P companies.

In February 2020, LendingClub announced it was acquiring Radius Bancorp, an internet bank in the U.S., LendingClub stated that the acquisition would provide benefits including lower funding costs, enhanced regulatory clarity, and business and revenue diversification. However, instead of interpreting the acquisition as LendingClub's desire to use its banking license to boost its P2P business, investors interpreted it as a compromise of the P2P giant's business model with banks after years of losses and lack of expansion. In October 2020, LendingClub announced that it would shut down its P2P platform by the end of 2020 and make its fourth change in 2021 – transforming into a full-service fintech marketplace bank. This further confirms that LendingClub acquired Radius Bancorp to get a banking license for its transformation and not to develop its P2P business further.

4.1.1 The LendingClub business model

As an internet-based platform, LendingClub operates entirely online with no traditional branch infrastructure. In terms of revenue composition, transaction fees are LendingClub's primary source of revenue, with that revenue coming primarily from transaction fees paid to LendingClub by card-issuing banks, education, and healthcare providers. The LendingClub platform plays a facilitating role in marketing to borrowers, and it helps issuing banks generate loans. Therefore, banks and education and healthcare providers pay transaction fees to LendingClub. The amount of these fees is based on the terms of the loan, including amount, grade, term and channel. Investor fees are another major source of revenue. These are the cost incurred by making loans, and they include managing borrower payments, collections, fees paid to investors, maintaining investor account portfolios, providing information, and publishing monthly statements. Investor fee income earned is primarily affected by the service fee rates paid by investors, the outstanding principal balance of loans and the amount of principal and interest collected from borrowers and remitted to investors.

As an information intermediary, LendingClub connects with borrower members and investor members.

I. Borrowing process

LendingClub provides unsecured personal loans for refinancing credit card balances and secured personal loans, mainly to refinance auto loans. Borrowers are required to register their accounts in their real names with the platform and provide supporting information before applying for a loan. To qualify for a loan, many factors are considered, including but not limited to: FICO score, satisfactory debt-to-income ratio, satisfactory credit history, and a limited number of credit inquiries within the past 6 months. When it receives a loan application, LendingClub provides the applicant with a variety of loan options, including the loan term, interest rate and amount for which the applicant qualifies. Once the applicant has selected a personalised financing option and completed the application process, LendingClub may conduct additional verification of the applicant. When the verification is complete, information about the borrower whose application is approved is posted on the platform's website. This includes the interest rate, term, total amount borrowed, and risk level, so investors can choose it. When the borrower and investor are successfully matched, WebBank disburses the loan to the borrower within a certain period and sells the proof of loan to LendingClub, thus allowing the investor to pay the funds to LendingClub. When the borrower repays the loan, LendingClub deducts a portion of the amount as a service fee and pays the remaining funds to the investor. If insufficient investor commitments are received and LendingClub does not choose to purchase the loan with its own funds, then the loan is unfunded.

II. Investing process

LendingClub first requires investors to register a real-name account, and the platform verifies the investor's information during registration. The investor's bank account is then linked to the account limit on the LendingClub platform through an automated clearing centre. Earnings generated from investments are transferred directly to the investor's bank account. LendingClub classifies borrowers into 35 levels (A1~G5) from low risk to high risk, based on a credit risk calculation model. A1-level borrowers have the lowest risk but the lowest corresponding interest rate; G5-level borrowers have the highest credit risk and the highest corresponding borrowing interest rate. Investors can choose to invest according to their risk preference.

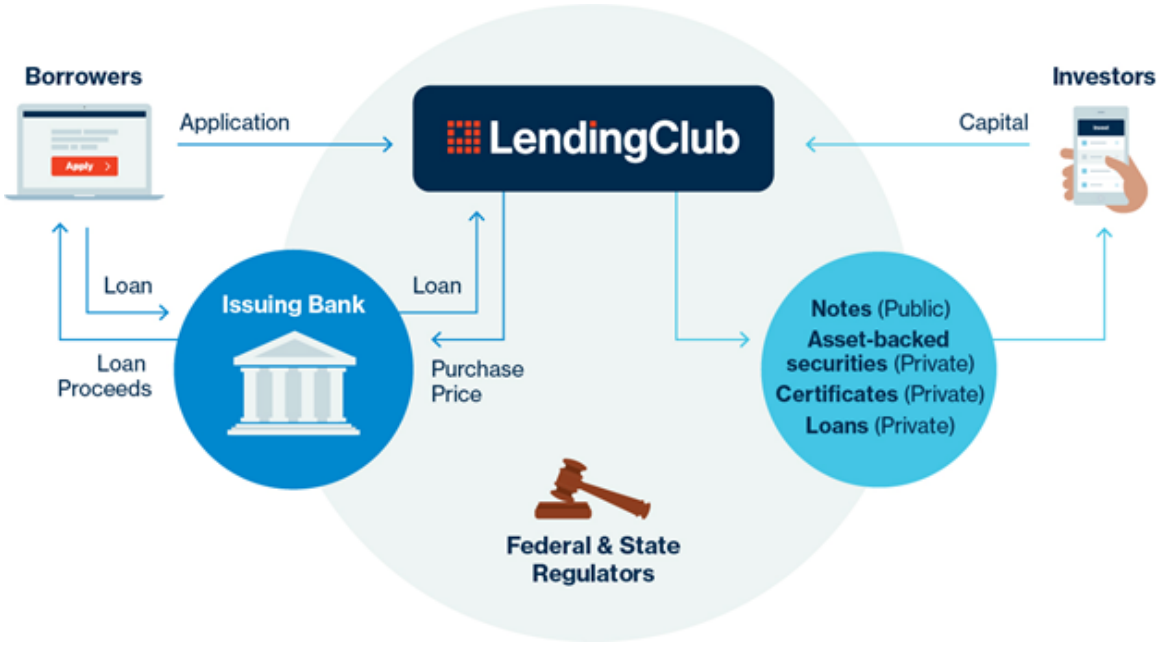

Figure 4.1: LendingClub business model. Source: Company 10-K

The data were collected from loans application evaluated by LendingClub between 2007 and 2018 (www.lendingclub.com), and the dataset was downloaded from Kaggle (www.kaggle.com). The data contain 2,260,701 observations with 151 variables each. The names of all the variables and the corresponding variable descriptions are shown in the appendix (Appendix Table 1), and Figure 4.2 shows the real data in the data analysis tool R.

Filter | Cols: « < 0 – 50 > »

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title
1	68407277	NA	3600	3600	3600	36 months	13.99	123.03	C	C4	leadman
2	68355089	NA	24700	24700	24700	36 months	11.99	820.28	C	C1	Engineer
3	68341763	NA	20000	20000	20000	60 months	10.78	432.66	B	B4	truck driver
4	66310712	NA	35000	35000	35000	60 months	14.85	829.90	C	C5	Information Systems Officer
5	68476807	NA	10400	10400	10400	60 months	22.45	289.91	F	F1	Contract Specialist
6	68426831	NA	11950	11950	11950	36 months	13.44	405.18	C	C3	Veterinary Tecnician
7	68476668	NA	20000	20000	20000	36 months	9.17	637.58	B	B2	Vice President of Recruiting
8	67275481	NA	20000	20000	20000	36 months	8.49	631.26	B	B1	road driver
9	68466926	NA	10000	10000	10000	36 months	6.49	306.45	A	A2	SERVICE MANAGER
10	68616873	NA	8000	8000	8000	36 months	11.48	263.74	B	B5	Vendor liaison
11	68356421	NA	22400	22400	22400	60 months	12.88	508.30	C	C2	Executive Director
12	68426545	NA	16000	16000	16000	60 months	12.88	363.07	C	C2	Senior Structural Designer
13	68338832	NA	1400	1400	1400	36 months	12.88	47.10	C	C2	Logistics Manager
14	66624733	NA	18000	18000	18000	60 months	19.48	471.70	E	E2	Software Manager
15	68466961	NA	28000	28000	28000	36 months	6.49	858.05	A	A2	Senior Manager

emp_length	home_ownership	annual_inc	verification_status	issue_d	loan_status	pymnt_plan	url
10+ years	MORTGAGE	55000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	65000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	63000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	110000	Source Verified	Dec-2015	Current	n	https://lendingclub.com/browse/loanDetail.action?lo...
3 years	MORTGAGE	104433	Source Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
4 years	RENT	34000	Source Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	180000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	85000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
6 years	RENT	85000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	42000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
6 years	MORTGAGE	95000	Not Verified	Dec-2015	Current	n	https://lendingclub.com/browse/loanDetail.action?lo...
1 year	MORTGAGE	70000	Not Verified	Dec-2015	Current	n	https://lendingclub.com/browse/loanDetail.action?lo...
3 years	MORTGAGE	64000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...
7 years	RENT	150000	Not Verified	Dec-2015	Charged Off	n	https://lendingclub.com/browse/loanDetail.action?lo...
10+ years	MORTGAGE	92000	Not Verified	Dec-2015	Fully Paid	n	https://lendingclub.com/browse/loanDetail.action?lo...

Figure 4.2 real data in R for part of 151 variables for the first 15 of 2,260,701 samples

4.2 Data Pre-processing

Real-world data can contain many missing values, a lot of noise, and outliers because of manual input errors, all of which can make it difficult to train algorithmic models. Data cleansing takes all kinds of dirty data and transforms them into standard, clean, continuous data for use in data statistics, data mining and other applications. The time spent on work related to data processing can be more than 70% of a project.

The dataset has 151 variables. To reduce the high-dimensional data to a more compact form – to avoid unnecessary computational stress and getting into overfitting problems – the dataset is first pre-processed. This section describes the steps of the pre-processing phase.

4.2.1 P2P Lending Loan Status Analysis

The ‘loan_status’ implies the current standing of the loan; it is the target variable for modelling. As the aim of modelling is to predict whether a user will default on a loan, this paper classifies the loans into two categories, good loans and non-performing loans, based on the loan_status.

In the original data, the variable ‘loan_status’ has 10 statuses. Their number and meaning are shown in Table 4.1.

Table 4.1 Descriptions of the target variable ‘loan_status’

Loan Status	Meaning	Observations

Fully Paid	Loan has been fully repaid, either at the expiration of the 3- or 5-year year term or as a result of a prepayment.	1076751
Current	Loan is up to date on all outstanding payments.	878317
Charged Off	Loan for which there is no longer a reasonable expectation of further payments.	268559
Late (31-120 days)	Loan has not been current for 31 to 120 days.	21467
In Grace Period	Loan is past due but within the 15-day grace period.	8436
Late (16-30 days)	Loan has not been current for 16 to 30 days.	4349
Does not meet the credit policy. Status: Fully Paid	While the loan was paid off, the loan application today would no longer meet the credit policy and wouldn't be approved on to the marketplace	1988
Does not meet the credit policy. Status: Charged Off	While the loan was charged off, the loan application today would no longer meet the credit policy and wouldn't be approved on to the marketplace	761
Default	Loan has not been current for 121 days or more.	40
NA	Null	33
Total		2260701

Previous studies have used a variety of classification strategies to categorise these 10 states, yielding somewhat different results. For this paper, default means being more than 30 days past due. Therefore, borrowers with a loan status of ‘late (30-120 days)’, ‘default’, and ‘charged off’ are classified as defaulters and labelled as ‘non-performing’, whereas borrowers with a loan status of ‘fully paid’, ‘in grace period’, and ‘late (16-30 days)’ are said to be in ‘good’. Loans that are not completed have no value to this paper, so samples with a ‘current’ status were removed, and because ‘NaN’ is a missing attribute

with no current status, those loans were removed. 'Does not meet the credit policy' means that loan application today would not be approved for the market. Such loans have no reference value, and they were removed.

The sample distribution after classification is shown in Figure 4.3.

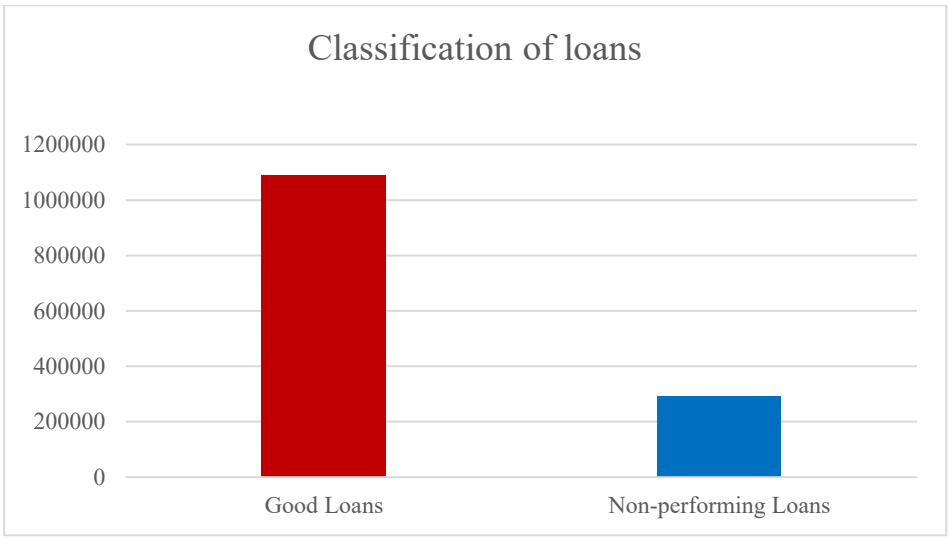

Figure 4.3 Classification of loans

With 1,089,536 good loans and 290,066 non-performing loans by classification, the good sample was 3.75 times larger than the non-performing sample. This is within the acceptable range for data mining modelling, as realistic defaulters represent only a small portion of overall users. However, we up-sample the groups in a later step to better extract signals that could lead to loan defaults.

4.2.2 Data Cleaning

I. Treatment of the missing value

The original dataset obtained from LendingClub contains a proportion of variables with missing values, which is common in real-world data. There is a variety of explanations for missing data, including device failures that result in entry failures, human reasons such as people hiding data for their own reasons, and different selection requirements that result in missing data in the early stages of the process.

This paper uses the rejection and imputation approaches to deal with missing values, depending on the type of variable and the proportion of missing values. Variables with a high proportion of missing values are rejected outright, as filling in values for such

variables may cause the model to deviate significantly from reality. There is no consistent cut-off in the literature for missing data in a dataset that is generally acceptable for statistical inference (Dong & Peng, 2013). Therefore, the threshold set here is based on the missing rate distribution (Figure 4-4).

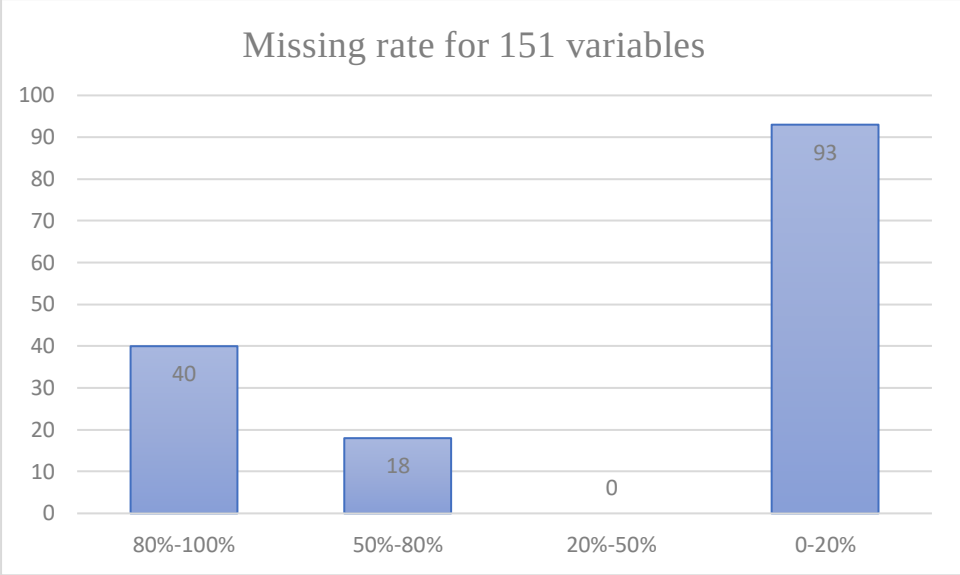

Figure 4.4 Missing rate for 151 variables

There were no variables in the dataset with missing rates between 20% and 50%, so the threshold was set at 20%. This means that variables with more than 80% missing data were removed. In this process, it is important to keep as much information as possible, but variables missing too much value may affect the model result even if we impute them. Null values in the variables were dealt with in the thesis using imputation techniques for variables with a smaller proportion of missing values. Categorical missing data were replaced with 'None', while numerical missing data were replaced with the median. Since less than 10% of the total feature data was replaced, and data on at least 40,000 loans were available for each variable, the imputation approach should have an insignificant effect on the study.

II. Treatment of invalid variables and those with low information value

Values that were not statistically meaningful were omitted. Some variables in the raw data are unique identifiers provided by the website to distinguish between different users. Thus, they are not useful in determining the borrower's credit danger. For example, 'URL' is the URL of the LendingClub page with the listing info, and 'ID' is the loan

listing's uniquely assigned ID. Because these are not needed for assessing the credit risk of a borrower, they have been removed. Variables that have only one attribute with zero variability, such as 'policy code', were removed.

In addition, categorical variables with an excessive number of unique values were omitted. These categorical variables contain a wealth of information. For example, 'Emp title' and 'title' have 365,752 and 59,997 unique values, respectively, but they are not useful for modelling and they increase the computational load. So, they have been removed. Besides, they are filled in manually by the borrower.

III. Feature derivation

This is the final pre-processing stage. For example, the correlation between 'fico_low_range' and 'fico_high_range' is almost 1. This means that they are highly correlated. So, after averaging, a new variable is derived and named 'fico_score'. Following the same process, a new variable 'last_fico_score' was created. The correlations of the other variables are determined later, as logistic regression is susceptible to multicollinearity.

$$'fico\_score' = ('fico\_high\_range' + 'fico\_low\_range') / 2 \quad (4)$$

$$'last\_fico\_score' = ('last\_fico\_high\_range' + 'last\_fico\_low\_range') / 2 \quad (5)$$

Based on the initial data cleaning procedures, 92 variables were omitted from the original dataset, leaving 59 variables, including 'Loan_status', as a dependent variable.

4.2.3 Data Normalisation

This is a pre-processing technique that is often used in machine learning. Scaling the features normalises the range of independent variables. Methods of feature scaling include min-max normalisation, mean normalisation, z-score normalisation and scaling to unit length. The values of different variables in the original dataset may differ significantly, and there is no comparability between variables since numerical variables also have different magnitudes. Variables with higher values become more useful in the analysis, and they affect the results if the raw data set is used. It is important to remove the effects of size differences and the range of values between numerical variables to consider problems in a systematic and integrated manner.

Z-score normalisation is used in this study to better deal with this problem. It allows for easier data comparison, and it reduces data redundancy. Data for numerical variables are first normalised by standard deviation in this article, with a mean of 0 and a standard deviation of 1. The transformation is as follows:

$$x_i = \big(x_i - mean(x)\big)/(std(x)) \tag{6}$$

4.2.4 Imbalanced Data Handling

Imbalanced data is a popular machine learning problem in which the total number of instances is less important than the set of instances in a category. Models trained on balanced datasets tend to outperform models trained on unbalanced data in terms of prediction. When a training dataset is imbalanced, the trained models are not robust and there is a large variation in the evaluation of the prediction results because of the large variation in the number of samples in different categories. When the number of samples in a category is sparse, it is very easy to ignore such samples during the training phase. So, the model ignores these features, reducing its prediction accuracy for such samples. Although the default rate of borrowers in P2P lending is higher than in traditional lending, the proportion of defaulters is still low. Moreover, there is a gap between on-time performers and late defaulters, so credit risk analysis of P2P lending must address that imbalance in the data set.

The distribution of information should be as even as possible. If there is a significant data imbalance, predictions are likely to be biased, i.e., classification results are skewed towards a larger group of observations. Sampling methods and manually produced data samples are often used to deal with imbalanced data. Oversampling from a few classes and undersampling from most classes are examples of sampling methods. Oversampling involves taking several samples from a minority class to increase the number of such samples in the training set to balance with the number of majority class samples. It is a fair increase in the number of minority class samples. Undersampling aims to minimise the sample size of several groups to get a balanced sample, i.e., to reduce the majority sample proportionately. With undersampling, eliminating the majority class sample may cause the model to lose much valuable knowledge about the majority class, while the simple replication in oversampling may lead to overfitting in model training. The SMOTE algorithm allows for better resolution of imbalanced data.

Chawla introduced the SMOTE algorithm in 2020 to address the issue of imbalanced data, and this thesis makes use of the SMOTE algorithm. This algorithm employs a technique called synthetic minority oversampling, which is an improvement over the random oversampling algorithm. The fundamental principle is to evaluate and model a small number of category samples first, and then add new artificially simulated samples to the data set so that the categories of the original data are no longer severely imbalanced (Figure 4.5) . The k-NN technique is used to simulate the algorithm's simulation process, and the following steps are taken to generate new samples:

- Sample the nearest neighbor algorithm to calculate k nearest neighbors for each minority class sample;
- Randomly select N samples from the k nearest neighbors for random linear interpolation;
- Construct a new minority class sample;
- Synthesise the new samples with the original data to produce a new training set.

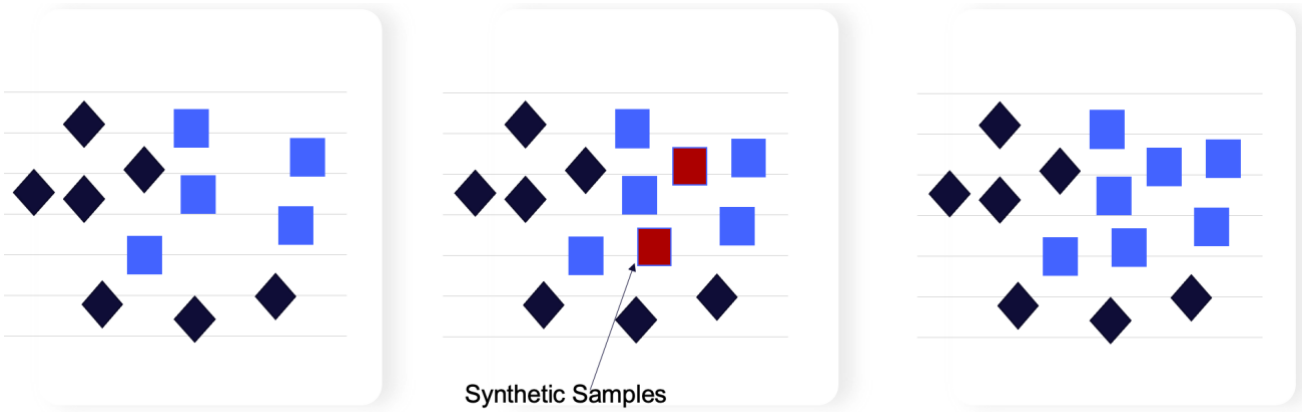

Figure 4.5 Synthetic Samples for SMOTE

4.2.5 Feature Selection

After the initial cleaning of the data, removing the loan status as the dependent variable left 57 variables. To analyse data, it is not the case that the more independent variables that are selected, the better the model results will be. Instead, additional input variables add extra degrees of freedom to the model itself. These help the model remember certain details, but they are not useful for building a stable, generalised model. That is, adding more uncorrelated variables increases the risk of over-fitting. Variable screening is conducted before machine learning to identify important variables that apply to the

dependent variable, prevent dimensional disasters, reduce the dimensionality of the data and further process the reduced data to improve the model.

I. Feature importance

According to Chen et al. (2020), the random forest varImp( ) method provides the highest accuracy and kappa for machine learning results when selecting main features for data classification, outperforming Boruta and recursive feature elimination (RFE). In this paper, the random forest method is used for the variable selection process. Before running the random forest method, one-hot encoding is applied for categorical variables. A higher importance score for a feature indicates that the variable has a greater influence on the dependent variable. The results show that the variable 'last_fico_score' is far more important than the other variables. However, to avoid dimensional problems and to include as much information about the borrower as possible, the top 20 variables were selected by comparing the importance scores of each variable (Figure 4.6).

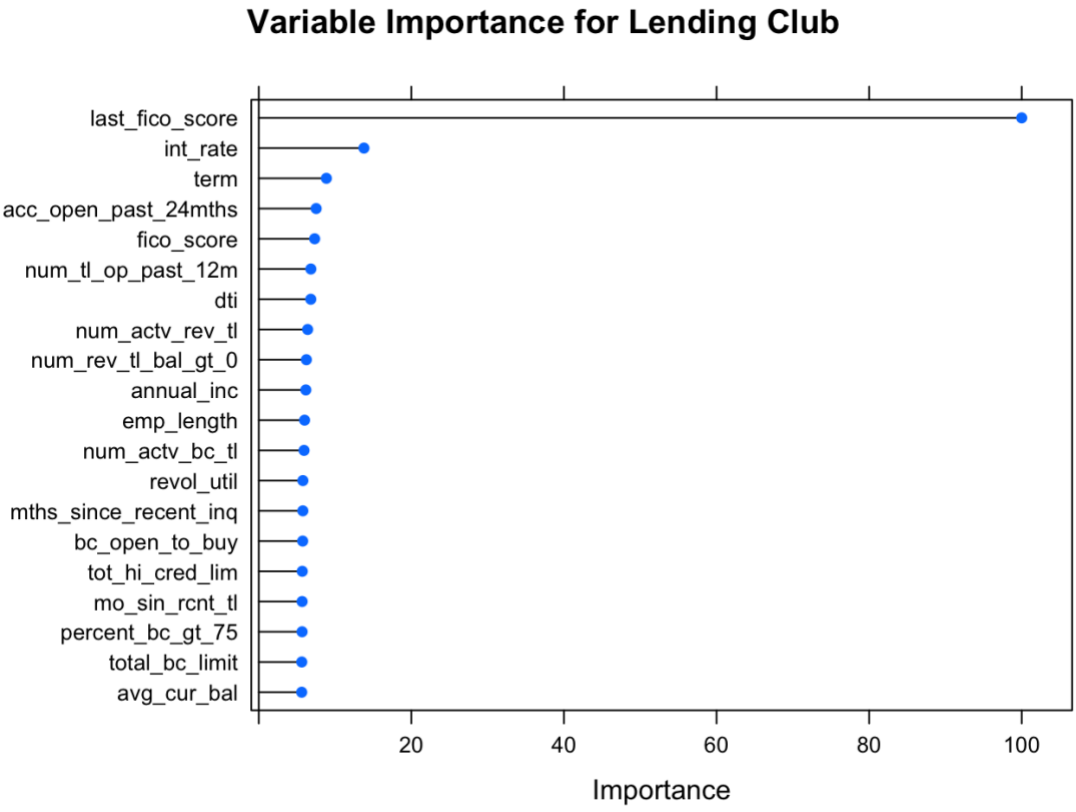

Figure 4.6 Variable importance for LendingClub

II. Autocorrelation

The autocorrelation between variables must be considered in logistic regression models. When the variables entering the model are highly correlated, the output of the model will be

affected and the variance of the regression coefficients will increase. The more variance variables have, the more difficult it is to interpret the coefficients. Therefore, this paper uses Pearson Correlation (Figure 4.7) and correlation matrix (Appendix Table 2) to check the correlation of the variables. For any two variables with a linear correlation coefficient above 0.7, the variable with the higher feature importance score is retained and the variable with the lower score is omitted. This measure resulted in the elimination of five variables and the retention of 15 variables (Table 4.2).

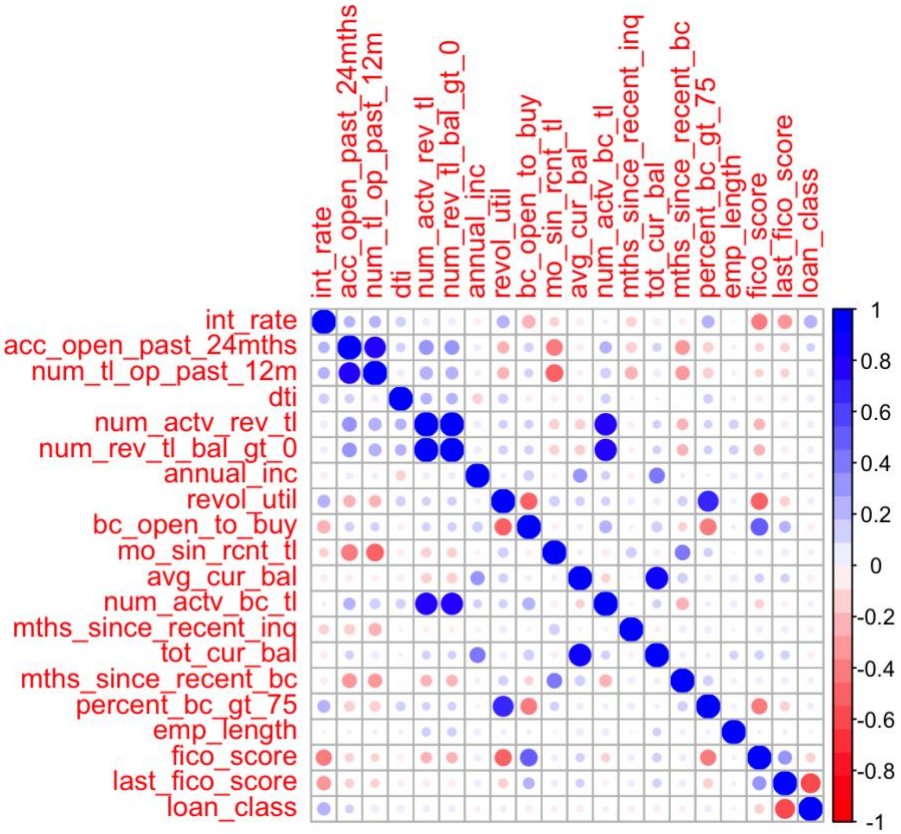

Figure 4.7 Pearson correlation for variables

Table 4.2 Data dictionary of final attributes utilised in our model

Variable Name	Description
last_fico_score	The borrower's last FICO Score
int_rate	Interest Rate on the loan
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
acc_open_past_24mths	Number of trades opened in past 24 months.

fico_score	The borrower's FICO score at loan origination
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
num_actv_rev_tl	Number of currently active revolving trades
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.
annual_inc	The self-reported annual income provided by the borrower during registration.
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
bc_open_to_buy	Total open to buy on revolving bankcards.
mo_sin_rcnt_tl	Months since most recent account opened
avg_cur_bal	Average current balance of all accounts
mths_since_recent_inq	Months since most recent inquiry.
mths_since_recent_bc	Months since most recent bankcard delinquency

4.3 Model Results & Discussion

In this section, we compare the performance of the four models with various benchmarks to provide a more well-rounded analysis of each model's performance. We present the models' accuracy, sensitivity, specificity, precision and F1 scores. The results of classification problems are often presented in confusion matrixes, which clearly indicate the performance in different categories. Below is an example of a confusion matrix.

Table 4.3 Confusion Matrix

	Reference: Good Loan	Reference: Bad Loan
Prediction: Good Loan	True Positive (TP)	False Positive (FP)
Prediction: Bad Loan	False Negative (FN)	True Negative (TN)

'Positive' and 'negative' describe the classes of the classification problem – in this case, the two loan statuses. In the code, we defined 'Good' to be positive and 'Non-performing' to be negative. 'True' and 'false' refer to the correctness of the prediction compared to the reference, which is the realistic scenario. 'True' means that our model predicts the result correctly, while 'False' indicates that the model makes a wrong prediction. In our case, 'False' could mean that the model predicts a non-performing loan to be a good loan or a good loan to be non-performing. In the next section, we provide a more detailed explanation of the benchmarks we chose for the confusion matrix.

Accuracy refers to the percentage of correct predictions relative to the entire sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Accuracy is both understandable and intuitive. However, it is not sufficient to only consider accuracy when comparing the performance of different models, especially under the scenario of imbalanced data. Overall accuracy is important when analysing model performance. However, due to the business implications of predicting defaults, we also care about the composition of accuracy. The composition refers to how many good loans are classified as good loans and how many non-performing loans are classified as non-performing. The prediction accuracy of non-performing loans is particularly relevant to this study because we would like to mitigate the default risk for P2P loans.

Precision refers to the percentage of samples that are positive among all samples that are predicted to be positive. While accuracy represents the overall prediction accuracy, including both positive samples and negative samples, precision represents the correctness of the prediction in the positive sample results.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Sensitivity is the probability that a positive sample will be predicted to be positive; in our case, it is the probability that good loans will be predicted as such.

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

The F1 score helps find a balance between precision and recall. The higher the F1 score, the better the performance of the model.

$$F1 = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{10}$$

Specificity refers to the percentage of good loans that are predicted to be good compared to all the good loans. This benchmark helps to measure whether we can identify all the good loans.

$$Specificity = \frac{TN}{FP + TN} \tag{11}$$

When comparing such benchmarks, the most favourable scenario is that we predict good loans to be 'good' and bad loans to be 'non-performing', which would mean that we make correct predictions. However, among incorrect predictions, false positives (FPs) are worse than false negatives (FNs). In FPs, we predict bad loans to be good. If investors invest in such loans, they will suffer a loss, and this will hurt the performance and reputation of our lending platform. In FN cases, although we miss good investment opportunities, we do not cause a loss of funds for investors. If we take a closer look at the functions above, we can note that the precision and the sensitivity functions look similar in format. The difference is between FPs and FNs as a part of the denominator. As discussed above, if we are to have a wrong prediction, we consider FNs better than FPs. A higher number of FNs means a larger number in the denominator, leading to a lower value for sensitivity. Therefore, when we compare the above benchmarks, we would prefer a prediction with lower sensitivity compared to precision, all else being equal.

Below are the confusion matrixes of all four models. We will discuss the results of each model in turn.

Table 4.4.1 Confusion Matrix for Logistic Regression

Logistic Regression-Confusion Matrix		
Prediction	Reference	
	good	non-performing
good	6306	415
non-performing	311	1872

Table 4.4.2 Confusion Matrix for Logistic Regression as Percentage

Logistic Regression-Confusion Matrix as Percentage		
Prediction	Reference	
	good	non-performing
good	0.9530	0.1815
non-performing	0.0470	0.8185

In the confusion matrix of the logistic regression model, the sum of actual good loans is 6617, and the sum of actual non-performing loans is 2287. The model wrongly predicts 311 good loans to be non-performing loans and 415 non-performing loans to be good loans, achieving an accuracy of 0.9185. In the correct predictions, the model predicts 95.30% of good loans and 81.85% of non-performing loans correctly. The model is better at predicting good loans correctly comparing with non-performing loans. In the wrong predictions, the model predicts 4.70% of good loans as non-performing loans and 18.15% of non-performing loans as good loans. The model identifies more non-performing loans to be good ones. While this model achieves the highest accuracy, it predicts more non-performing loans as good loans than good loans as non-performing.

Table 4.5.1 Confusion Matrix for Random Forest

Random Forest-Confusion Matrix		
Prediction	Reference	
	good	non-performing
good	6279	395
non-performing	338	1892

Table 4.5.2 Confusion Matrix for Random Forest as Percentage

Random Forest-Confusion Matrix as Percentage		
Prediction	Reference	
	good	non-performing
good	0.9489	0.1727
non-performing	0.0511	0.8273

In the confusion matrix of the random forest model, the sum of actual good loans is 6617, and the sum of actual non-performing loans is 2287. The model wrongly predicts 338 good loans to be non-performing loans and 395 non-performing loans to be good loans, achieving an accuracy of 0.9177. In the correct predictions, the model predicts 94.89% of good loans and 82.73% of non-performing loans correctly. The model is better at predicting good loans correctly comparing with non-performing loans. In the wrong predictions, the model predicts 5.11% of good loans as non-performing loans and 17.27% of non-performing loans as good ones. The model identifies more non-performing loans to be good ones. This model achieves the second-highest prediction accuracy, but it predicts more non-performing loans as good ones than good loans as non-performing, like the logistic regression model.

Table 4.6.1 Confusion Matrix for k-NN

KNN-Confusion Matrix		
Prediction	Reference	
	good	non-performing
good	5879	659
non-performing	738	1628

Table 4.6.2 Confusion Matrix for k-NN as Percentage

KNN-Confusion Matrix as Percentage		
Prediction	Reference	
	good	non-performing
good	0.8885	0.2882
non-performing	0.1115	0.7118

In the confusion matrix of the k-NN model, the sum of actual good loans is 6617, and the sum of actual non-performing loans is 2287. The model wrongly predicts 738 good loans to be non-performing loans and 659 non-performing loans to be good loans, achieving an accuracy of 0.8431. In the correct predictions, the model predicts 88.85% of good loans and 71.18% of non-performing loans correctly. The model is better at predicting good loans correctly comparing with non-performing loans. In the wrong predictions, the model predicts 11.15% of good loans as non-performing loans and 28.82% of non-performing loans as good ones. The model identifies more non-performing loans to be good ones. This model achieves the lowest prediction accuracy, and it predicts the largest percentage of non-performing loans as good ones than good loans as non-performing.

Table 4.7.1 Confusion Matrix for SVM

SVM-Confusion Matrix		
Prediction	Reference	
	good	non-performing
good	6334	475
non-performing	283	1812

Table 4.7.2 Confusion Matrix for SVM as Percentage

SVM-Confusion Matrix as Percentage		
Prediction	Reference	
	good	non-performing
good	0.9572	0.2077
non-performing	0.0428	0.7923

In the confusion matrix of the SVM model, the sum of actual good loans is 6617, and the sum of actual non-performing loans is 2287. The model wrongly predicts 283 good loans to be non-performing loans and 475 non-performing loans to be good loans, achieving an accuracy of 0.9149. In the correct predictions, the model predicts 95.72% of good loans and 79.23% of non-performing loans correctly. The model is better at predicting good loans correctly comparing with non-performing loans. In the wrong predictions, the model predicts 4.28% of

good loans as non-performing loans and 20.77% of non-performing loans as good ones. The model identifies more non-performing loans to be good ones. This model achieves the third-highest prediction accuracy, but it predicts more non-performing loans as good ones than good loans as non-performing, like the logistic regression model.

Table 4.8 Matric Summary for Four Models

	Logistic Regression	Random Forest	KNN	SVM
Accuracy	0.9185	0.9177	0.8431	0.9149
Sensitivity	0.9530	0.9489	0.8885	0.9572
Specificity	0.8185	0.8273	0.7118	0.7923
Precision	0.9383	0.9408	0.8992	0.9302
F1 Score	0.9456	0.9448	0.8938	0.9435

To further compare the four models, we present the accuracy, sensitivity, specificity, precision and F1 score in all four models in one chart and rank the benchmarks between each model. The logistic regression model has the highest accuracy and F1 score. It also has the second-highest sensitivity, specificity, and precision. Hence, we consider the logistic regression model to provide the best prediction among the four models. The random forest model has the second-highest accuracy and F1 score. It has the highest specificity and precision, and the third-highest sensitivity. Therefore, we consider it as the second-best-performing model. The SVM model has the highest sensitivity, but all other benchmarks are ranked third; we consider this model to have the third-best performance. The k-NN model performs relatively poorly compared to the other three; it has the lowest rankings in all four benchmarks, making it the worst-performing model.

Table 4.9 Ranking Summary for Four Models

Ranking for Four Machine Learning Models				
	Logistic Regression	Random Forest	KNN	SVM
Accuracy	1	2	4	3
Sensitivity	2	3	4	1
Specificity	2	1	4	3
Precision	2	1	4	3
F1 Score	1	2	4	3
Total	1	2	4	3

Chapter 5 Conclusion

The conclusion section will include a review of the research question, a summary of outer applications based on this paper, limitation, and recommendations for future research.

The following research question was to be answered in this thesis:

- *Among the selected machine learning methods, which method performs best in default prediction in peer-to-peer lending for a given model evaluation metric?*

In summary, the results in Chapter 4 show that the logistic regression model is more adaptable and accurate at predicting borrower credit risk for P2P lending. It is therefore more appropriate for the systematic consideration of borrower default risk, which is also supported by theoretical validation. The logistic regression model, random forest model, SVM, and k-NN were ranked according to their effectiveness in predicting the credit risk of P2P borrowers.

From researching and summarising the P2P lending industry, we can see the advantage of providing people with lower credit record access to financing. Reportlinker.com (2020) announced in its 'Global Peer-to-peer Lending Industry' report that the global P2P lending market size will grow at a CAGR of 42.7% over the analysis period 2020–2027. By offering a better predictive model of credit risk, we can improve the reputation of the online P2P lending platform and help the platform gain a better position in the fast-growing market.

This paper aims to compare different methods for predicting credit defaults using the data of LendingClub loans. After careful research to obtain the required data, we conducted data cleaning and feature selection through multiple dimensions. During these processes, we considered both financial and statistical interpretability. We experimented with four machine learning algorithms: logistic regression, random forest, support vector machines (SVMs) and k-nearest neighbours.

The contribution of this paper is that we expand the data to a 10-year horizon and implement multi-dimensional design in data preparation, balancing financial and statistical interpretability. We use a comprehensive combination of benchmarks to evaluate the performance of each model. We found that the best-performing model is the random forest. The ranking from best to worst is logistic regression, random forest, SVM, and KNN. The k-NN model performs worst among all four models because it has the lowest sensitivity and the lowest precision. This paper demonstrates the utility of comparing the performance of machine learning models in a well-rounded manner because different benchmarks provide different information.

This paper can serve as a reference for outer applications of P2P lending platforms and has three main implications. While conventional risk management relies on data such as basic consumer information and credit history to assess creditworthiness, this paper starts from 151 variables that describe the features of borrowers from a high-dimensional perspective. We employ four widely used machine learning techniques that are promising according to previous studies to forecast defaults in P2P lending. As can be shown, P2P lending platforms that incorporate additional aspects of information can help mitigate defaults, even without collateral. Therefore, the first implication is that high-dimensional data should be considered in risk management activities, including default recognition. The second implication is that a careful and reasonable data preparation process can improve the accuracy of prediction. Future studies could test different theories to balance between financial and statistical interpretability. The third implication is that this paper evaluates the performance of machine learning models in a more well-rounded way than simply comparing their accuracy. Moreover, we start with the confusion matrix to develop a series of benchmarks to compare the models' performance; this could also be beneficial for future studies.

This paper shows that achieving accurate identification of default risk in P2P lending can help protect the rights of investors and reduce losses due to default. For the purposes of regulating the P2P lending industry, models that can accurately identify P2P lending defaults represent the risk management capabilities of a P2P platform. A mature P2P network should have a risk management framework that leverages vast user data and data mining techniques to mitigate risk and maintain a consistent level of profitability. When performing modelling work, P2P platforms can maximise data's value for default identification by fully using the available data mining technologies. Platforms can use mature classical data mining

techniques, such as logistic regression, support vector machines and the other models discussed in this paper, that have been developed over time and have a strong theoretical foundation. Such techniques will ensure the models' predictability and business interpretability. In future research, P2P platforms can experiment with more cutting-edge data mining techniques.

This paper lays the foundation for future studies that evaluate investment strategies. Compared to a random investment strategy in which loans are randomly selected for investment, a default-based investment strategy can improve investment returns to a certain extent. Based on this paper, returns on default-based investment strategies may be studied further for research or business purposes.

In future research, we could quantify the economic gains of improving the lending model with machine learning algorithms. The business value of this process is that it could be used for financial planning in internal control and for marketing purposes to illustrate the competence of the platform. However, calculating the return could be a rather complicated process, as it depends on the loan amount recovered and the total time horizon of loan collection. For example, it is possible for a lender to recover a portion of a loan during the loan repayment process, even if the loan is in default. With regard to the time frame, although LendingClub loans have a term of 36 or 60 months, borrowers can repay their loans early and early recovery may increase the return on the loan but may also lead to reinvestment issues.

Reference

Abdou, H., Pointon, J. and El-Masry, A. (2008) 'Neural nets versus conventional techniques in credit scoring in Egyptian banking', *Expert Systems with Applications*, 35(3), pp. 1275–1292. doi: 10.1016/j.eswa.2007.08.030.

Barasinska, N. and Schäfer, D. (2014) 'Is Crowdfunding Different? Evidence on the Relation between Gender and Funding Success from a German Peer-to-Peer Lending Platform', *German Economic Review*, 15(4), pp. 436–452. doi: https://doi.org/10.1111/geer.12052.

Bertsch, C. and Rosenvinge, C.-J. (2019) *FinTech credit: Online lending platforms in Sweden and beyond*. doi: 10.13140/RG.2.2.25272.08965.

Brennan, P-J. (2009). Peer-to-peer lending lures investors with 12% return (update 2). Bloomberg.

Byanjankar, A., Heikkilä, M. and Mezei, J. (2015) 'Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach', *2015 IEEE Symposium Series on Computational Intelligence*, pp. 719–725.

Jbs.cam.ac.uk. 2018. Cambridge Centre for Alternative Finance, *The 3rd Americas Alternative Finance Industry Report.* [online] Available at: <https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/2019-05-ccaf-3rd-americas-alternative-finance-industry-report.pdf> [Accessed 17 May 2021].

Chen, D., Li, X. and Lai, F. (2017) 'Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China', *Electronic Commerce Research*, 17(4), pp. 553–583. doi: 10.1007/s10660-016-9247-2.

Chen, R.-C. *et al.* (2020) 'Selecting critical features for data classification based on machine learning methods', *Journal of Big Data*, 7(1), p. 52. doi: 10.1186/s40537-020-00327-4.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20(3), pp. 273–297. doi: 10.1007/BF00994018.

Desai, V. S., Crook, J. N. and Overstreet, G. A. (1996) 'A comparison of neural networks and linear scoring models in the credit union environment', *European Journal of Operational Research*, 95(1), pp. 24–37. doi: 10.1016/0377-2217(95)00246-4.

Dong, Y., HAO, X. and SATO, H. (2015) 'Investigation of the Impact of Data Comparability on Performance of Support Vector Machine Models for Credit Scoring', *Innovation and Supply Chain Management*, 9, pp. 31–38. doi: 10.14327/iscm.9.31.

Dong, Y. and Peng, C.-Y. J. (2013) 'Principled missing data methods for researchers', *SpringerPlus*, 2(1), p. 222. doi: 10.1186/2193-1801-2-222.

Eisenbeis, R. A. (1978) 'Problems in applying discriminant analysis in credit scoring models', *Journal of Banking & Finance*, 2(3), pp. 205–219. doi: https://doi.org/10.1016/0378-4266(78)90012-2.

FitzPatrick, P. J. . P. D. (1932) ‘A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies’, *Journal of Accounting Research*.

Freedman, S. and Jin, G. (2008) ‘Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper.Com’, *NET Institute,* Working Papers. doi: 10.2139/ssrn.1304138.

Galindo, J. and Tamayo, P. (2000) ‘Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications’, *Computational Economics*, 15(1), pp. 107–143. doi: 10.1023/A:1008699112516.

Gonzalez, L. and Komarova, Y. (2014) ‘When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans’, *Journal of Behavioral and Experimental Finance*, 2, pp. 44–58. doi: 10.1016/j.jbef.2014.04.002.

Henley, W. E. and j. Hand, D. (1997) ‘Construction of a k-nearest-neighbour credit-scoring system †’, *IMA Journal of Management Mathematics*, 8(4), pp. 305–321. doi: 10.1093/imaman/8.4.305.

Herzenstein, M. *et al.* (2008) ‘The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-to-peer Lending Communities’, in papers.ssrn.com.

Huang, C., Chen, M. and Wang, C.-J. (2007) ‘Credit scoring with a data mining approach based on support vector machines’, *Expert Syst. Appl.*, 33, pp. 847–856.

Imtiaz, S. and Brimicombe, A. (2017) ‘A Better Comparison Summary of Credit Scoring Classification’, *International Journal of Advanced Computer Science and Applications*, 8. doi: 10.14569/IJACSA.2017.080701.

Iyer, R. *et al.* (2009) ‘Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending?’, *Harvard University, John F. Kennedy School of Government, Working Paper Series*. doi: 10.2139/ssrn.1570115.

James H. Myers and Edward W. Forgy (1963) ‘The Development of Numerical Credit Evaluation Systems’.

Jin, Y. and Zhu, Y. (2015) ‘A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending’, *2015 Fifth International Conference on Communication Systems and Network Technologies*, pp. 609–613.

Khandani, A., Kim, A. and Lo, A. (2010) ‘Consumer Credit-Risk Models Via Machine-Learning Algorithms’, *Journal of Banking & Finance*, 34, pp. 2767–2787. doi: 10.1016/j.jbankfin.2010.06.001.

Klafft, M. (2008) ‘Peer to Peer Lending: Auctioning Microcredits over the Internet’.

Larrimore, L. *et al.* (2011) ‘Peer to Peer Lending: The Relationship Between Language Features, Trustworthiness, and Persuasion Success’, *Journal of Applied Communication Research*, 39. doi: 10.1080/00909882.2010.536844.

Lenz, R. (2016) 'Peer-to-Peer Lending: Opportunities and Risks', *European Journal of Risk Regulation*, 7, pp. 688–700. doi: 10.1017/S1867299X00010126.

LendingClub (2020) Corporation Company 10-K

Malekipirbazari, M. and Aksakalli, V. (2015) 'Risk Assessment in Social Lending via Random Forests', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2015.02.001.

Pope, D. and Sydnor, J. (2008) 'What's in a Picture? Evidence of Discrimination from Prosper.com', *Monetary Economics eJournal*, 46. doi: 10.1353/jhr.2011.0025.

Ravina, E. (2007) 'Beauty, Personal Characteristics, and Trust in Credit Markets', *SSRN Electronic Journal*. doi: 10.2139/ssrn.972801.

Ravina, E. *et al.* (2012) *Love & Loans The Effect of Beauty and Personal Characteristics in Credit Markets 1 for helpful comments and suggestions*.

Research, M., Services, F. and Trends, B., 2021. *Global Peer-to-peer Lending Industry*. [online] Reportlinker.com. Available at: <https://www.reportlinker.com/p05900042/Global-Peer-to-peer-Lending-Industry.html?utm_source=GNW> [Accessed 17 May 2021].

Steenackers, A. and Goovaerts, M. J. (1989) 'A credit scoring model for personal loans', *Insurance: Mathematics and Economics*, 8(1), pp. 31–34. doi: https://doi.org/10.1016/0167-6687(89)90044-9.

Tang, H (2018): "Peer-to-peer lenders versus banks: substitutes or complements?", Review of Financial Studies, forthcoming.

Van Gestel, T. *et al.* (2003) 'A support vector machine approach to credit scoring', *Bank en Financiewezen*, 2, pp. 73–82.

WangDaiZhiJia, 2019. *China P2P Lending: Number of Platform.* [online] Ceicdata.com. Available at: <https://www.ceicdata.com/en/china/p2p-lending-number-of-platform> [ccessed 17 May 2021].

West, D. (2000) 'Neural network credit scoring models', *Computers & Operations Research*, 27(11), pp. 1131–1152. doi: https://doi.org/10.1016/S0305-0548(99)00149-5.

Wiginton, J. C. (1980) 'A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior', *The Journal of Financial and Quantitative Analysis*, 15(3), pp. 757–770. doi: 10.2307/2330408.

Yum, H., Lee, B. and Chae, M. (2012) 'From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms', *Electronic Commerce Research and Applications*, 11(5), pp. 469–483. doi: 10.1016/j.elerap.2012.05.003.

Zekic-Susac, M., Sarlija, N. and Benšić, M. (2004) *Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models*. doi: 10.1109/ITI.2004.241696.

Appendix

Table 1 Data Dictionaries

LoanStatNew	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened

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.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.
last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
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.
loan_status	Current status of the loan
max_bal_bc	Maximum current balance owed on all revolving accounts
member_id	A unique LC assigned Id for the borrower member.
mo_sin_old_il_acct	Months since oldest bank installment account opened

mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
mths_since_rcnt_il	Months since most recent installment accounts opened
mths_since_recent_bc	Months since most recent bankcard account opened.
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_inq	Months since most recent inquiry.
mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
next_pymnt_d	Next scheduled payment date
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
open_acc	The number of open credit lines in the borrower's credit file.
open_acc_6m	Number of open trades in last 6 months
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_act_il	Number of currently active installment trades
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors

pct_tl_nvr_dlq	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.
policy_code	publicly available policy_code=1 new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently in the borrower's credit file
total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment accounts
total_bc_limit	Total bankcard high credit/credit limit
total_cu_tl	Number of finance trades
total_il_high_credit_limit	Total installment high credit/credit limit
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
total_rev_hi_lim	Total revolving high credit/credit limit
url	URL for the LC page with listing data.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified

zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances
sec_app_fico_range_low	FICO range (high) for the secondary applicant
sec_app_fico_range_high	FICO range (low) for the secondary applicant
sec_app_earliest_cr_line	Earliest credit line at time of application for the secondary applicant
sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant
sec_app_open_acc	Number of open trades at time of application for the secondary applicant
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts
sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant
sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant
sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant
sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant
sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant
hardship_flag	Flags whether or not the borrower is on a hardship plan
hardship_type	Describes the hardship plan offering
hardship_reason	Describes the reason the hardship plan was offered
hardship_status	Describes if the hardship plan is active, pending, canceled, completed, or broken
deferral_term	Amount of months that the borrower is expected to pay less than the contractual monthly payment amount due to a hardship plan
hardship_amount	The interest payment that the borrower has committed to make each month while they are on a hardship plan
hardship_start_date	The start date of the hardship plan period
hardship_end_date	The end date of the hardship plan period
payment_plan_start_date	The day the first hardship plan payment is due. For example, if a borrower has a hardship plan period of 3 months, the start date is

	the start of the three-month period in which the borrower is allowed to make interest-only payments.
hardship_length	The number of months the borrower will make smaller payments than normally obligated due to a hardship plan
hardship_dpd	Account days past due as of the hardship plan start date
hardship_loan_status	Loan Status as of the hardship plan start date
orig_projected_additional_accrued_interest	The original projected additional interest amount that will accrue for the given hardship payment plan as of the Hardship Start Date. This field will be null if the borrower has broken their hardship payment plan.
hardship_payoff_balance_amount	The payoff balance amount as of the hardship plan start date
hardship_last_payment_amount	The last payment amount as of the hardship plan start date
disbursement_method	The method by which the borrower receives their loan.
debt_settlement_flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.
debt_settlement_flag_date	The most recent date that the Debt_Settlement_Flag has been set
settlement_status	The status of the borrower's settlement plan.
settlement_date	The date that the borrower agrees to the settlement plan
settlement_amount	The loan amount that the borrower has agreed to settle for
settlement_percentage	The settlement amount as a percentage of the payoff balance amount on the loan
settlement_term	The number of months that the borrower will be on the settlement plan

Table 2 Correlation Matrix

	int_rate	acc_open_past_24mths	num_tl_op_past_12m	dti	num_actv_rev_tl	num_rev_tl_bal_gt_0	annual_inc	revol_util	bc_open_to_buy
int_rate	1	0.188200144	0.203028402	0.146380519	0.085132373	0.084877612	-0.071220924	0.239919055	-0.271460487
acc_open_past_24mths	0.188200144	1	0.754053873	0.119486243	0.33767157	0.329863532	0.057129211	-0.219158174	0.119002874
num_tl_op_past_12m	0.203028402	0.754053873	1	0.069312506	0.258150914	0.2453111	0.052231047	-0.211393594	0.10631111
dti	0.146380519	0.119486243	0.069312506	1	0.190035029	0.191970455	-0.139315153	0.139210107	-0.052364625
num_actv_rev_tl	0.085132373	0.33767157	0.258150914	0.190035029	1	0.982006057	0.07306789	0.103956224	0.104218497
num_rev_tl_bal_gt_0	0.084877612	0.329863532	0.2453111	0.191970455	0.982006057	1	0.071920575	0.112773897	0.115553455
annual_inc	-0.071220924	0.057129211	0.052231047	-0.139315153	0.07306789	0.071920575	1	0.035671933	0.156053478
revol_util	0.239919055	-0.219158174	-0.211393594	0.139210107	0.103956224	0.112773897	0.035671933	1	-0.46212443
bc_open_to_buy	-0.271460487	0.119002874	0.10631111	-0.052364625	0.104218497	0.115553455	0.156053478	-0.46212443	1
mo_sin_rcnt_tl	-0.121707021	-0.434073486	-0.49327205	-0.065827636	-0.14483497	-0.140829528	-0.02407555	0.155035242	-0.050490569
avg_cur_bal	-0.078843489	-0.073078488	-0.047930045	-0.076054392	-0.151886819	-0.153524838	0.301529509	0.122214023	0.045106281
num_actv_bc_tl	0.024371513	0.221297205	0.161725834	0.129540558	0.808613478	0.801549003	0.104078648	0.107005228	0.240321996
mths_since_recent_inq	-0.14472477	-0.169742001	-0.248591793	0.007058002	-0.052791574	-0.043860643	-0.035132686	0.08042066	-0.015958618
tot_cur_bal	-0.079384846	0.101030164	0.086678605	0.012933071	0.096770818	0.09620004	0.389918409	0.081089962	0.167407547
mths_since_recent_bc	-0.080518561	-0.331429642	-0.301072796	-0.000508191	-0.194407226	-0.191544944	0.028178416	0.135979168	-0.107652596
percent_bc_gt_75	0.24334836	-0.162314511	-0.171210566	0.126757894	0.09408966	0.100817062	-0.015506222	0.704906804	-0.438849268
emp_length	-0.004971263	0.023237234	0.025332661	0.02143693	0.118978575	0.119763582	0.064782874	0.034754522	0.026159746
fico_score	-0.402502826	-0.102600803	-0.091330029	-0.06009188	-0.185229245	-0.184001645	0.069406083	-0.458981604	0.486936455
last_fico_score	-0.312178551	-0.116667013	-0.104378591	-0.058428248	-0.080257087	-0.079317862	0.064791689	-0.135835558	0.196148606

	mo_sin_rcnt_tl	avg_cur_bal	num_actv_bc_tl	mths_since_recent_inq	tot_cur_bal	mths_since_recent_bc	percent_bc_gt_75	emp_length	fico_score	last_fico_score
int_rate	-0.121707021	-0.078843489	0.024371513	-0.14472477	-0.079384846	-0.080518561	0.24334836	-0.004971263	-0.402502826	-0.312178551
acc_open_past_24mths	-0.434073486	-0.073078488	0.221297205	-0.169742001	0.101030164	-0.331429642	-0.162314511	0.023237234	-0.102600803	-0.116667013
num_tl_op_past_12m	-0.49327205	-0.047930045	0.161725834	-0.248591793	0.086678605	-0.301072796	-0.171210566	0.025332661	-0.091330029	-0.104378591
dti	-0.065827636	-0.076054392	0.129540558	0.007058002	0.012933071	-0.000508191	0.126757894	0.02143693	-0.06009188	-0.058428248
num_actv_rev_tl	-0.14483497	-0.151886819	0.808613478	-0.052791574	0.096770818	-0.194407226	0.09408966	0.118978575	-0.185229245	-0.080257087
num_rev_tl_bal_gt_0	-0.140829528	-0.153524838	0.801549003	-0.043860643	0.09620004	-0.191544944	0.100817062	0.119763582	-0.184001645	-0.079317862
annual_inc	-0.02407555	0.301529509	0.104078648	-0.035132686	0.389918409	0.028178416	-0.015506222	0.064782874	0.069406083	0.064791689
revol_util	0.155035242	0.122214023	0.107005228	0.08042066	0.081089962	0.135979168	0.704906804	0.034754522	-0.458981604	-0.135835558
bc_open_to_buy	-0.050490569	0.045106281	0.240321996	-0.015958618	0.167407547	-0.107652596	-0.438849268	0.026159746	0.486936455	0.196148606
mo_sin_rcnt_tl	1	0.037263172	-0.0785731	0.167534017	-0.058623496	0.36798248	0.114273453	-0.00629939	0.059920743	0.069634332
avg_cur_bal	0.037263172	1	-0.110245248	-0.005890531	0.835905126	0.153082552	0.0568887	0.086723721	0.1109018	0.104969474
num_actv_bc_tl	-0.0785731	-0.110245248	1	-0.0216452	0.104898143	-0.2366626	0.034436427	0.074266448	-0.112689443	-0.041337674
mths_since_recent_inq	0.167534017	-0.005890531	-0.0216452	1	-0.040334414	0.071469353	0.071418982	0.003386536	0.054032006	0.061083043
tot_cur_bal	-0.058623496	0.835905126	0.104898143	-0.040334414	1	0.053354173	0.027807043	0.09903939	0.119084137	0.108742369
mths_since_recent_bc	0.36798248	0.153082552	-0.2366626	0.071469353	0.053354173	1	0.148437623	0.037951593	0.064908568	0.074752715
percent_bc_gt_75	0.114273453	0.0568887	0.034436427	0.071418982	0.027807043	0.148437623	1	0.025486545	-0.397899383	-0.136631928
emp_length	-0.00629939	0.086723721	0.074266448	0.003386536	0.09903939	0.037951593	0.025486545	1	0.017446699	0.023721101
fico_score	0.059920743	0.1109018	-0.112689443	0.054032006	0.119084137	0.064908568	-0.397899383	0.017446699	1	0.294421655
last_fico_score	0.069634332	0.104969474	-0.041337674	0.061083043	0.108742369	0.074752715	-0.136631928	0.023721101	0.294421655	1