Tilburg University

School of Economics and Management

pwc

Master Thesis Finance

# PEER-TO-PEER LENDING AND PREDICTING DEFAULT OF BORROWERS

Analyzing the characteristics of peer-to-peer (P2P) lending and forecasting the default rate of a P2P platform from the United States

Berend Johannes Henricus Maria van den Boomen

Student number: 1266999

b.j.h.m.vdnboomen@tilburguniversity.edu

Internship at PwC

Supervisor Tilburg University: Fabio Braggion

Supervisor PwC: Bert Hartholt

- 2020 –

# Table of Contents

# Abstract

This paper investigates the financial characteristics of a peer-to-peer (P2P) lending platform (LendingClub) and analyses if the dependent variable default can be predicted. First, borrower- and loan features on loan default are explored to provide insight into the relationship between these characteristics and default risk. Borrower characteristics as debt-to-income, income, FICO score, and homeownership are significantly influencing the default rate. On the other hand, loan features as the applied amount, interest rate, and loan maturity also have an adequate impact on the forecast of default. Second, a logistic regression is performed to forecast the probability of a defaulted loan. Lastly, a ROC curve is estimated to examine the accuracy rate of default. The ROC curve compares different variables and different models to establish the most accurate model. Overall, the established model is, with an accuracy ratio of 70.33% against an accuracy ratio of 63.56%, better in predicting default than the model of LendingClub.

**Keywords**: P2P lending, default, predicting, logistic regression, investment strategy, ROC curve

# Preface

This piece of paper is one of the most valuable papers of my fantastic student life. With this piece of paper, my education journey at Tilburg University and Instituto Superior de Economia e Gestao (Lisboa) will come to an end with a Master of Science degree in Finance.

First of all, I particularly wish to express my gratitude to my coach at PwC Bert Hartholt for his support, discussions, comments, and advice while writing this thesis at PwC, Amsterdam. Unfortunately, due to the COVID-19 virus, I was not always able to come to the office, but he and PwC, in general, were very generous with helping me writing my thesis and first work experience. In September 2020, I will start my professional career at PwC Amsterdam, which makes me excited and proud. Additionally, I want to thank prof. Fabio Braggion for the useful feedback, introduction in P2P lending, and for giving me the right direction. The whole process of writing my Master thesis and studying in the field of Finance has increased my personal, analytical, and research skills.

Secondly, my appreciation goes towards my friends at fraternity K.O.N.G.S.I., and fellow students during the Master Finance for the stimulating discussions, for all the coffee and lunch breaks together. More importantly, for all the fun we have had in the last year, despite the extraordinary COVID-19 virus.

Lastly, my gratefulness towards my parents (Tom and Inge), godfather (Jan), and grandparents (Jan, Joke, Harry and Maria) that allowed me to experience this incredibly educational and social journey. Finally, my love and appreciation for all the support of my sisters (Anouk, Femke and Babette), and, especially, my girlfriend, Arlette.

# 1. Introduction

Since the financial crisis of 2008, the barriers of credit access have increased in the United States (U.S.) and Europe (Calabrese, Osmetti, & Zanin, 2019). These barriers have increased the development and diffusion dispersal of alternative financial services to constitutional banking for credit access, such as peer-to-peer (P2P) lending platforms (known as the FinTech credit market).

FinTech is the development of technology within the financial world. FinTech includes every company which uses the Internet, connects with mobile devices, and software technology or cloud services to execute or connect with financial institutions[1]. FinTech has the potential to create financial services to worldwide consumers currently lacking access and to implement a new way of financing money. Financial technology covers digital and technical innovations in the financial sector to improve and automate the use of financial services. Whereas traditional banking institutions receive interest, P2P lending platforms generate income by charging service fees and commissions for using the platform. The benefits of reducing costs are applicable for both borrowers and lenders. From the lender's perspective, removing the traditional bank reduces overhead costs from investing and from the borrower's perspective, search- or switching costs can be shortened (Balyuk, 2016).

New technological innovations are essential to the finance industry since it has been captivating billions of dollars in venture capital over the last decade (Guild, 2017). According to Jagtiani and Lemieux (2018), the personal unsecured loan market in the U.S. had reached nearly $112 billion in the third quarter of 2017. Whereas the FinTech investors only had 3 percent of this market in 2010, these investors are now projected to have 30 percent market share. Other contributors to the market are constitutional banks, finance businesses, and insurance companies. Despite the rise in consumer credit over the last decade, Bricker et al. (2017) mention that 20.8 percent of the consumer households felt credit limited, and this outcome has been stable in the last years on the report of the Survey of Consumer Finance in 2016.

Furthermore, the European Commission approved an action plan on FinTech in March 2018 to develop the financial industry with more competitive and innovative features, which indicates that this new methodology of financing needs to be investigated. The plan supports new technologies such as Blockchain, Artificial Intelligence, P2P lending and the increase of the integrity of the financial system (European Commission, 2018).

In contrast, the main problem with P2P lending is that there is information asymmetry between the negotiating parties (Yum, Lee, & Chae, 2012). Predicting the default of borrowers and controlling the borrower is a critical element of the lending process. Within the FinTech platforms,

[1] https://www.thestreet.com/technology/what-is-fintech-14885154

forecasting the creditworthiness of borrowers is difficult and, therefore, asking for the right characteristics and using the convenient calculating mechanisms is essential for lenders or Dutch banks.

Despite the studies on P2P lending and new FinTech techniques, there is little research conducted on the prediction of default and the risk within P2P platforms. If the indicators of negligence would be investigated, investors have less chance of losing their money. This study examines the characteristics and the forecast of default within a US P2P lending platform, namely LendingClub.

There is still much to investigate regarding P2P lending and the incorporated models. Therefore, this thesis will examine different hypothesis to answer the research question and provide more insight into the new development within the financial world. Several studies (Jagtiani & Lemieux, 2018; Michels, 2012; Iyer, Khwaja, Luttmer, & Shue, 2009) mention significant features of P2P lending and this indicates that characteristics decide whether to provide a loan or not. To better understand the P2P online lending and the used models, this paper will examine the factors of online credit and tries to model the default rate. Therefore, the conducted research question is as follows: *What are the financial determinants of default within P2P lending, and can we model the default rate?*

Moreover, there will be three hypotheses examined to interpret the research question clearer. These are: *(1): Which financial characteristics have an impact on the default risk?, (2): Can the default rate be forecasted based on these financial determinants?, and (3) Can a computer-based algorithm outperform the system of LendingClub in terms of default prediction?*

This thesis is different from previous literature since recent research has not focused on prognosticating default rate within P2P online lending. This research will enhance the existing literature by using data from LendingClub (a U.S. founded P2P platform) between 2009 and 2016 to answer the research question and provide insights into the foreseeing of default in the FinTech industry.

This dissertation is structured as follows. Chapter 1 provides an introduction to P2P lending and states the research question. In chapter 2, an extensive overview of the literature regarding P2P lending is discussed. Chapter 3 explains the sample selection and describes the dataset investigated in this paper. Subsequently, chapter 4 ascertains the methodology used to test the hypotheses. Chapter 5 lists and analyses the empirical results. Finally, chapter 6 formulates the discussion, conclusion, and limitations and recommendations for future research.

# 2. Literature review

The section will describe a review of the relevant literature. In subsection 2.1, an introduction to peer-to-peer lending (P2P) will be provided. Subsection 2.2 illustrates the financial determinants of a P2P loan.

## 2.1 Peer-to-Peer (P2P) Lending

### 2.1.1 Definition

Peer-to-peer (P2P) lending describes the loan origination between private individuals (peers) on online platforms where financial institutions operate only as intermediates required by law (Bachmann et al., 2011). Initialized by groups in online social networks, the first commercial online P2P lending platform Zopa started in 2005. P2P lending creates the opportunity to connect lenders and borrowers through the Internet and engage in loan transactions. This new type of credit is one of the significant innovations in which FinTech competes with traditional banks. Furthermore, the borrowers of P2P platforms consist of riskier applicants which result in higher default rates (de Roure, Pelizzon, & Thakor, 2018). Hazardous debtors make the borrowers of LendingClub less interesting for traditional banks and will increase costs.

P2P lending is most common in the consumer credit market. FinTech lending differs from traditional banks in two significant measures. First, FinTech uses algorithms to automate the loan process and demand zero or minimum human work (Fuster, Plosser, Schnabl, & Vickery, 2019). Second, FinTech does not have capital requirement constraints issued by deposit-taking businesses.

FinTech has been playing an increasing player in the financial and banking world. There have been considerations about the implementation of other data sources by FinTech providers and the impact on economic incorporation. According to Jagtiani and Lemieux (2019), there is a high correlation with interest rates, LendingClub (ranking) rates and loan performance. For the same risk of default, consumers have fewer spreads on loans from LendingClub than from credit card borrowing. As mentioned by Jagtiani and Lemieux (2018), the quality of the information provided for P2P is crucial but different to justify. Researchers have examined identifying information that could be leveraged in an online money claim. Michels (2012) discovered that voluntary disclosure of hard income as income, income source, education level, and other debt decreases the interest rate. Gao and Lin (2013) analyzed different requests and used text mining to find that more complex descriptions correlate with higher default rates.

Furthermore, Hassin and Trope (2000) show that people form expectations that are heavily forced and potentially biased by appearance assumptions. Hassin and Trope (2000) examine whether these appearance expectations turn into financial actions and model social outcomes, for example, elections. Figure 1 shows an analysis of whether differences in the trustworthiness of borrowers are related to differences in funding success and interest rates.

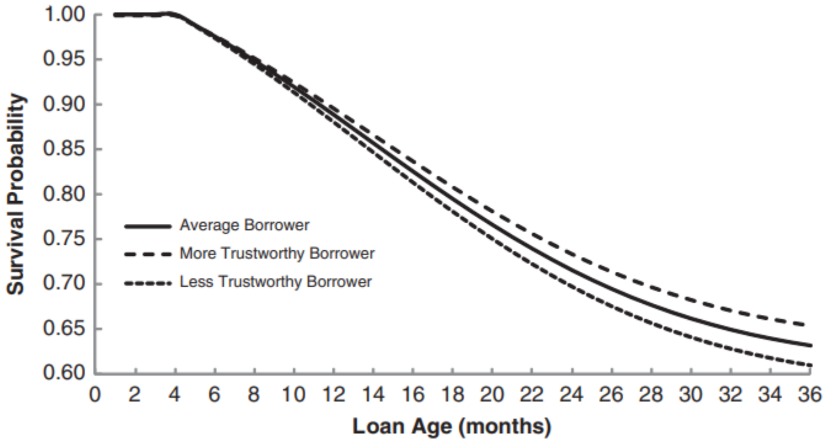


*Figure 1: Relative loan survival by different borrowers over loan maturity. Adopted from Hassin and Trope (2000).*

Consistent with the trust-intensive nature of lending, Figure 1 illustrates that borrowers who appear more trustworthy, in terms of impressions of attractiveness or financial resources, have higher possibilities of receiving their requested loan (Duarte, Siegel, and Young, 2012). Borrowers who borrower more often are indeed more trustworthy, have better credit scores and have a lower risk of default. To conclude, the level of trustworthiness affects financial agreements, because this level has a negative relation with the default risk and decreases the interest rate.

### 2.1.2 Moral hazard and adverse selection

Traditional financial institutions approve credit only after an extensive check of the borrowers' characteristics and the resulting score from credit risk models. This practice is widely applied by risk management to address (at least in part) the matter of moral hazard and adverse selection due to information asymmetry. Like traditional credit channels, lenders in the P2P market can take advantage of credit scoring forecasts to discriminate between potential good and bad borrowers. Such predictions are obtained by applying statistical models and machine learning algorithms trained on a sample of past borrowers for which several socio-economic and demographic characteristics and behavioural factors are known.

Moral hazard is the risk that a stakeholder in the agreement has not entered into a contract in good faith or has given incorrect information about its assets, liabilities, or credit capacity. Also, moral hazard may mean a participant has an incentive to take extraordinary risks in an anxious attempt to receive an income before the contract ends. Moral hazard is an essential issue within P2P lending since the borrower can provide misleading or misleading information and the borrower can ask for different interest rates when they are aware of their (to less) financial income.

Besides, adverse selection is another problem that comes with P2P lending through an online platform. Adverse selection refers to a situation in which borrowers have information that investors do not have, or vice versa, about some aspect of borrower quality. In that case, there is asymmetric information, which means that one party has more excellent material knowledge than the other party. Adverse selection and moral hazard are two serious problems with online lending and, therefore, needs to be investigated and analyzed.

### 2.1.3 Previous research

Previous research on P2P lending is new and relatively small. There are different exciting subjects in which the study can be divided. First, the qualification to provide a P2P loan is essential for lenders to conduct (Balyuk, 2016). Credit intermediaries are dependent on the classification of P2P lenders in the decision to provide loans to borrowers, which makes it essential which aspects determine the possibility of credit.

Secondly, there is research constructed on the sources of default in P2P lending (Larrimore, Jiang, Larrimore, Markowitz, & Gorski, 2011). Larrimore et al. (2011) mention, for example, that the use of textual description and quantitative words to justify the loan is related to the proportion of successful investments. The use of language in online P2P lending is one of the characteristics that have an impact on the amount and possibility of a loan.

Thirdly, other research that has been done is the screening ability of lenders in the P2P community. Decisive factors in determining whether someone can repay his loan is conducted in the following researches (Iyer, Khwaja, Luttmer, & Shue, 2016; Duarte, Siegel, & Young, 2012). Duarte et al. (2012) find that borrowers who apply more often for loans have higher probabilities of getting their loans funded. Furthermore, civilizing personal data for financial situations are negatively correlated with receiving an investment on P2P platforms.

Lastly, previous research suggests that LendingClub loans increases in areas where the local economy is not performing well (Jagtiania and Lemieux 2018). Jagtiania and Lemieux (2018) find that LendingClub's consumer lending activities have incorporate areas that are underserved by traditional banks, typically in highly concentrated economies and countries where there are fewer bank facilities per person.

### 2.1.4 P2P lending and traditional banking

Online P2P platforms have increased over the last decade, and this was part of the unregulated banking system (Feng, Fan, & Yoon, 2015). There are no entry thresholds nor clear supervision rules to analyze this new type of financing. These financial intermediaries provide banking service but have uninsured deposits and are not accountable for capital requirements. They give an approximation of how financial institutions would be backed when they did not have subsidized deposits nor liable to capital requirements.

Klafft (2008) mentioned that the rules of traditional banking are almost identical to the guidelines for P2P lending. Klafft (2008) used another popular P2P platform in the U.S., namely Prosper, which show that the credit rating of borrowers is the most crucial variable in choosing the interest rate. In contrast, the debt to income ratio is less outstanding but still has significant importance.

Milne and Parboteeah (2016) mention that P2P lending is complementary to traditional banks, but not to be compared. Buchak, Matvos, Piskorski and Seru (2018) test what the impact of technology is on P2P platforms and the influence of new marketplace lenders in general. Buchak et al. (2018) conclude that P2P lenders are better able to supply more creditable customers than conventional banks. P2P loans need to be adopted by banks, either by cooperation between P2P platforms or providing proprietary platforms. The risks, business and regulatory issues in P2P lending needs to be taken into development, which includes communication, control of liquidity risks, minimalization of fraud, security and operational risks.

Still, as with every bank, there are anxieties about the regulations. Because of the increasing growth of the nonbanks, this implies some regulatory questions. FinTech platforms collect soft information about trustworthiness in a nontraditional way, primarily due to algorithms and computer-based software. This outcome of these computers may be unfair and, therefore, provide incorrect information towards the investors (Jagtiani & Lemieux, 2018).

De Roure, Pelizzon, and Thakor (2018) exploit a model with endogenous options of a traditional bank against P2P lending and investigate their developments using a German database. The findings were that P2P loans have, on average, more risk than bank loans, whereas risk-adjusted rates on P2P loans are below the ones on bank loans. Also, the authors mention that P2P platforms are growing because some banks are challenged with exogenously more regulatory costs.

## 2.2 Financial determinants

One of the main problems in online P2P lending is information asymmetry. P2P platforms must reduce the principal-agent problem (Jensen & Meckling, 1979). The principal-agent problem arises when the lender wants to receive as much valid information about the borrower or deal as possible; the borrower might want to hide some of his components to receive a lower interest rate. P2P platforms tend to allow their borrowers to decide an investment built on accurate information and reliable financial information that has been validated by external parties. Subsequently, most platforms demand users to enhance demographic characteristics, such as family, age or birthplace.

P2P lending platforms arrange an overview of economic determinants of the borrower and lender as the primary index for creditworthiness. Typical financial features are credit ratings, monthly income and expenses, family and home status, and the debt to income ratio. These characteristics are generally reviewed by external rating companies that combine personal and financial features to a credit-score (Klafft, 2008).

Iyer et al. (2009) analyzed the effect of the borrower's credit rating on their funding success. The creditworthiness of borrowers from other characteristics than credit rating is asked as well. According to Iyer et al. (2009), 28 percent of the interest spread between the highest credit rating category (A.A.) and the least trustworthy borrower (HR), is determined by other characteristics than the credit rating (Figure 2). Other attributes in this paper are debt-to-income ratio, the number of current delinquencies or the number of credit inquiries. Non-standard variables have a fewer impact on the interest rate, corresponding to their research.

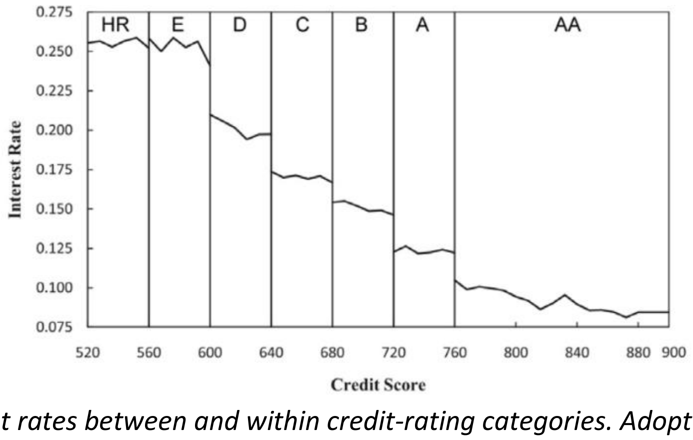


*Figure 2: Interest rates between and within credit-rating categories. Adopted from Iyer et al. (2009).*

Furthermore, Lin et al. (2017) discovered that educational level and working years have a significant effect on the default risk. Both working over a more extended period and higher education have a lower probability of default. As mentioned by Chen et al. (2019), both loan period and interest rates have a positive relation towards the default risk. Jin and Zhu (2015) mention that loan term, annual income, loan amount, *DTI*, and credit rate are important variables referring to the prediction of default.

### 2.2.1 Hard- and soft information

Besides financial determinants to clarify interest rate or default risk, there needs to be a mix of hard financial information and soft demographic factors. Additional tests should be performed to indicate the creditworthiness of a borrower and to reduce information asymmetries. Soft demographic characteristics imply non-financial information, like listing texts, place of birth and residence, family and sometimes pictures. An essential part of the review of data is the status of verifying (Michels 2012). This review is especially crucial since P2P online platforms tend to put less exertion in the screening of borrower's data than the conventional bank does.

Frequently, the online platforms verify the financial details like credit score, bank account information and credit history. Many studies find a significant part for though information to clarify default in P2P lending. For example, Serrano-Cinca et al. (2015) state that the factors of justifying default within LendingClub are credit grades, income, the purpose of the loan, family, and credit history. This outcome indicates that investors are offered information regarding the analyzation of borrower's creditworthiness. Moreover, Iyer et al. (2010) suggest that investors on Prosper infer the majority part of borrower's creditworthiness based on hard information. Overall, Iyer et al. (2010) conclude that the lenders can deduce future borrower default, but that this implication is insufficient.

However, the question of how the operators of P2P platforms manage soft information is more critical. As previously mentioned, P2P is emerging, and this may cause a loss of soft information due to long banking relationships. Dorfleitner et al. (2016) mention that soft information hardly prognosticates the default rate on two leading European platforms which are in Germany, namely Smava and Auxmoney. Dorfleitner et al. (2016) conclude that control variables as solvency scores and mainly interest rates are significant factors for the probability of default. They find that high-interest rates show a positive relation with default probability. Investors on P2P platforms are influenced by soft information when deciding upon investing. The difficult information (like credit score, income, and household) is solely dependent on providing the granted loan.

P2P platforms can identify people who would not receive a loan by a bank or elsewhere. Investors need to set an interest rate which is corresponding with the problematic information. Additionally, another exciting feature of P2P lending is the use of social networks. Borrowers can form partnerships with others or form community groups. However, this term is heterogeneous, because friends or community groups are, unfortunately, not always mentioned or detectable.

# 3. Sample selection and data description

The dataset of one of the leading United States FinTech platforms: LendingClub. LendingClub is a leading United States peer-to-peer lending platform, which publishes their data (freely) public for investors to evaluate their portfolio and approve their investment. The loan dataset of LendingClub includes (daily) updated loan data from all lenders, which is not protected by data-protection laws. The dataset is downloaded from their website on March 30th, 2020.

## 3.1 LendingClub

LendingClub is the platform on which the empirical analysis is established. LendingClub offers loans with a duration of 36- or 60-months. Further, this thesis limits the research with a timeframe from 2009 till 2016 because these are the most recent and reliable years. The loans before 2009 have limited availability of some explanatory variables, which can cause endogeneity in the regression. Within the final sample, this thesis focusses only on loans with a status of "Fully Paid", "Default" or "Charged Off". Loans are titled “Charged Off” when there is no longer a fair probability of future payments (after 120 days or more). Both "Late payment" and "In grace period" are removed, because tracking of costs goes beyond the capacity of this dissertation and these two statuses are neglectable because they are with a minimal amount. This process leaves a sample of 725,641 different loans.

Borrowers that apply for a loan must report the following data: name, address, the purpose of the requested loan and loan amount. The platform demands the applicant's identity to enhance information about the corresponding credit report and rating. Afterwards, LendingClub calculates the debt to income (*DTI*) ratio and reports the *FICO score*. The *FICO score* indicates the creditworthiness of a borrower and is requested to be over 660. When the borrower passes this screening test, LendingClub proposes several loans with different interest rates, maturities (either 36- or 60 months), and other sums. Subsequently, an applicant chooses a proposed investment from the options; the loan request is displayed on the website of LendingClub and becomes available to lenders. Potential investors, which can either institutions or persons, observe different loan characteristics and analyze details from the borrower's credit report. Temporarily, LendingClub requests the borrower to report his income, source of income, and the length of employment.

According to LendingClub, the Securities and Exchange Commission (SEC), in 2013, said that 79% of the applicants had their employment or income verified. To verify income, LendingClub requests documents such as recent paychecks, tax returns, or bank statements; to verify employment, LendingClub can contact the working place directly or refer to other data. There are several possibilities for LendingClub also to withdraw the request or not to fund the loan: (1) the loan was deleted grounded on "a credit decision or the inability to verify certain borrower information"; (2) the borrower withdrew himself the request; and (3) the loan amount was not fully funded. LendingClub mentioned that almost all listed loans are provided full investing, and most of them within a few days. The corresponding area is in a total of 46 states of the United States and the District of Columbia. During the sample period (2009 – 2016Q4), all loans were financed by individual investors.

## 3.2 The dataset

The dataset includes in total 725,641 loans during the period 2009 till 2016. An overview of the variables with the corresponding description is shown in Table 1. From this data, only loans issued between January 2009 and December 2016 were used for the analysis. The period before January 2009 is excluded because the financial crisis of 2008 might disrupt the dataset and significance of the variables. The exclusion of 2017 and onwards is because the maturity is not yet reached so that an analyzation would give misleading outcomes. While the time frame is indicated, there are loans which are determined to reach already their maturity and have not been fully paid back or defaulted. The excluded loans are generated by some late payment in the credit life cycle. Late instalment enlarges the full maturity of the credit.

Classification of the loans in the dataset is whether a loan is default (1) or paid (0). The statistical program Stata evaluates the probability of default and the analyzation of the dataset. Whether a loan is paid or default depends on the status. Paid loans are with the level of "fully paid", and default loans are with the status of "default" or "charged-off". Loan characteristics are the requested credit amount, the loan maturity and the interest rate. Borrower characteristics are income, income source verified, *FICO score*, age, LendingClub rating, education, employment length, debt-to-income and house.

Missing data is handled as follows. There are loans which include missing data, and these loans are still included in the model because Stata and R can handle missing data in their regressions. Therefore, the more extensive the dataset, the more precise the analyzation. Missing data can occur due to applicants who not fully provided all the requested information or information was not recorded.

## 3.3 Pricing of the loan

LendingClub provides a credit rating between 35 credit grades, which have a range from A1 to G5. This credit grade is calculated on the borrower's *FICO score*, *DTI* ratio, debt history, requested credit amount, and maturity. After the credit grade is selected, the interest rate will be evaluated and assigned. LendingClub operates with a national pricing policy, which includes that every state has the same *FICO score* and interest rate. This national pricing policy is besides consistent with the expectation that P2P credit supply is elastic, mentioned in the paper by Tang (2019). LendingClub operates with a two-step approach when calculating the interest rate for each loan: (1) assigning a credit rating and grade; and (2) calculate the interest rate within the platform's base rate plus an additional adjustment corresponding to the cited factors.

An interesting question arises when the regression will be tested against the credit grading of LendingClub. *Is it possible to be more accurate in terms of predicting default than the model of LendingClub?* LendingClub's model includes every variable in their calculation, whereas this thesis will only contain 10 (relevant) variables.

## 3.4 Descriptive statistics

The total dataset includes 150 different variables, but not all variables are relevant to form for the analyzation of default. Useful variables are shown in Table 2.1. Table 2.1 displays the descriptive statistics for all variables in the dataset. The variables are segregated in three groups, which are loan result, loan characteristics and borrower characteristics.

According to Table 2.1, the probability of *Default* within the LendingClub platform is 17.33%, whereby the average outstanding loan is 14,333 dollars. This moderate risk of default is relatively high, which is corresponding to the average of 20 percent stated in the paper of Croux, Jagtiani, Korivi, and Vulanovic (2020). LendingClub allows riskier borrowers to borrow at their platform, which results in higher default risk and, therefore, a high average number of charged off loans.

*DTI* has an average of 18.06%, meaning that the borrower spends 18.06% of his or her monthly income on debt. *DTI* is a measure of comparing the net income of borrowers and provides an extra indication in the wealth of a borrower. Other impressive results are that the average interest rate is 13.09%, and the average rating score is 4.34. *Rating* score is provided by LendingClub and has a range from A (1) to G (7). An average rating score of 4.34 means that the average *Rating* is D. This result is consistent with the fact that many borrowers have no income source verified (mean is equal to 2.043). No income verification indicates that these borrowers are not willing to provide their (financial) resources. LendingClub explains that it determines the grade based on credit information of the borrower and other provided data from the borrower. LendingClub's interest rate is calculated by LendingClub's base rate and an addition for risk and volatility.

Moreover, the educational level of the loan borrowers is relatively high, namely 3.9 out of 5. A reason for this might be that more educated people are aware of platforms like LendingClub, and are, therefore, more the type of user of LendingClub.

*FICO score* is constructed by Fair Isaac Corporation (FICO) and includes a number from <500 (very poor) to 800+ (exceptional) in terms of creditworthiness of borrowers[2]. *FICO score* gives an overview of the borrower's creditability. *FICO score* is applied by more than 90% of the major United States lenders[3].

Table 2.2 displays descriptive statistics for two different samples, namely where the default rate is measured (*Default = 1)* and where the loan has been repaid. Table 2.2 provides an insight into the different means, medians, standard deviations, minimum and maximum. Loans with a higher amount have more probability of default since the mean of *Default* is higher ($15,462) than the mean of repaid loans ($14,084). Additionally, *Default* loans were expected to be riskier and were determined to have higher interest rates, which is corresponding with Table 2.2. Interest rate, on average, have a mean of 15.5% for *Default* loans against a mean of 12.7% for loans which are paid off.

Furthermore, the *DTI* is higher for loans that are *Default* (19.85) than for repaid loans (17.67). This outcome indicates that the higher the *DTI*, the more risk on default for loans. *Annual income* and *FICO score* are also both higher for *Default*, respectively $76,785 and 699.8 against $69,653 and 689.8. These numbers are in line with the paper of Lin, Prabhala, and Viswanathan (2013). They discovered that requested loans are less likely to be financed and expected to default more when having a lower credit rating. The longer, on average, the loan, the more probability of default, since the mean term for *Default* is 45.91 months against 41.14 months for repaid loans. An explanation would be that longer maturity increases the possibility of financial instability of borrowers and thus a longer time frame to default.

As mentioned by Jagtiani, Lambie-Hanson, and Lambie-Hanson (2019), FinTech lenders provide more loans towards borrowers with lower credit scores and lower annual income. The reason for this is corresponding to the risk at these borrowers and the higher interest rates.

[2] https://www.myfico.com/credit-education/what-is-a-fico-score

[3] https://www.mybanktracker.com/credit-cards/credit-score/which-credit-score-the-most-accurate-264522

# 4. Methodology

This section describes the methodology for this thesis to answer the research questions. First, the experimental setup is reported in subsection 4.1. Secondly, the performance of LendingClub is analyzed and discussed in subsection 4.2.

## 4.1 Experimental setup

This thesis will use an empirical method from earlier research by Iyer et al. (2016) to evaluate the criteria for P2P lending. This paper considers a peer-to-peer market where lenders view both financial information and soft information about the borrower. To check whether the borrower will default a logistic regression will be conducted, and the accuracy rate of the regression will be the determinant if the model can predict default.

Firstly, this research will examine a screening performance to borrower quality, as replaced by ex-post loan performance. Secondly, this thesis investigates how lenders weight different sources of information informing their screening measure. Thirdly, this dissertation measures whether lenders can infer borrower creditworthiness as measured by ex-post default along dimensions not captured by the *FICO score*. Finally, the output (a ROC curve) will be investigated to determine the default rate of lenders.

To investigate whether the model forecasts the probability of default, a logit model and a linear probability model are conducted because this regression has a binary output (default or not default). The performance measurement is the accuracy rate of the computer to predict the default. This logistic regression is performed after the elements of the P2P lending are known and will be delivered with Stata and R. Because "default" is a binary outcome, standard ordinary least squares (OLS) assumptions such as homoscedasticity and linearity are violated. Consequently, a logistic regression model and a logit model with a binomial probability model are assumed to be more successful. The regression is established on a panel data estimator. The estimators have desirable properties only if critical assumptions as unbiasedness, efficiency, consistency and asymptotic normality hold. Standard errors are clustered at the borrower level because borrowers might apply for more than one loan.

A logistic regression forecasts the relationship between default risk and corresponding financial characteristics with the following model:

$$\begin{aligned}\text{Probability}(Default_{it} = 1) \\ &= \beta_0 + \beta_1 \log{(Applied\ Amount)_{it}} + \beta_2\ Loan\ Term_{it} + \beta_3\ Interest\ rate_{it} \\ &+ \beta_4 \log(Income)_{it} + \beta_5\ Verified_{it} + \beta_6\ Fico\ Score_{it} + \beta_7\ House_{it} + \beta_8\ DTI_{it} \\ &+ E_{it} + G_{it} + \theta_t\end{aligned}$$

Where *i* refers to the borrower, and *t* refers to calendar years. The borrower and loan characteristics included in the model are the loan amount requested by the borrower (*Applied Amount*), the duration of the loan (*Loan Term*), the interest rate of a borrower (*Interest rate*), the annual income (*Income*), a dummy variable for the verification of the income source (1 if verified and 0 otherwise) (*Verified)*, the *FICO score* (*FICO score*), a dummy variable for house ownership (1 for a mortgage, 0 otherwise) (*House*), the debt-to-income (*DTI*) ratio, a vector for the *length of Employment* ($E_{ijt}$) and a vector for *LendingClub's grade (*$G_{it}$*)*, and $\theta_t$ denotes time fixed effects.

Additionally, a logistic regression will be conducted to investigate if *DTI* ratios are more effective in preventing *Default* for low credit rating borrowers (with a *Grade* between D and G). This regression is as follows:

$$\begin{aligned}\text{Probability}(Default_{it} = 1) \\ &= \beta_0 + \beta_1\ Low\ Rating_{it} + \beta_2\ DTI_{it} + \beta_3\ Low\ Rating_{it} * DTI_{it} + \theta_t\end{aligned}$$

Where *i* refers to the borrower, and *t* refers to calendar years. The borrower and loan characteristics included in the model are a dummy variable for the low rating of a borrower (1 if Grade between D and G, and 0 if Grade is between A and C), the debt-to-income (*DTI*) ratio, and $\theta_t$ denotes time fixed effects.

Fixed effects are an estimator within panel data that loses time-invariant variables that can be correlated with the independent variable *Default*[4]. For example, when β is related to a variable, the regression cannot reliably approximate the vector of parameters of interest β using OLS since the underlying assumption of no correlation within error term and regressors are violated. In cross-section, strategies to correct this omitted variable issue are instrumental variables or the use of fixed effects in panel data.

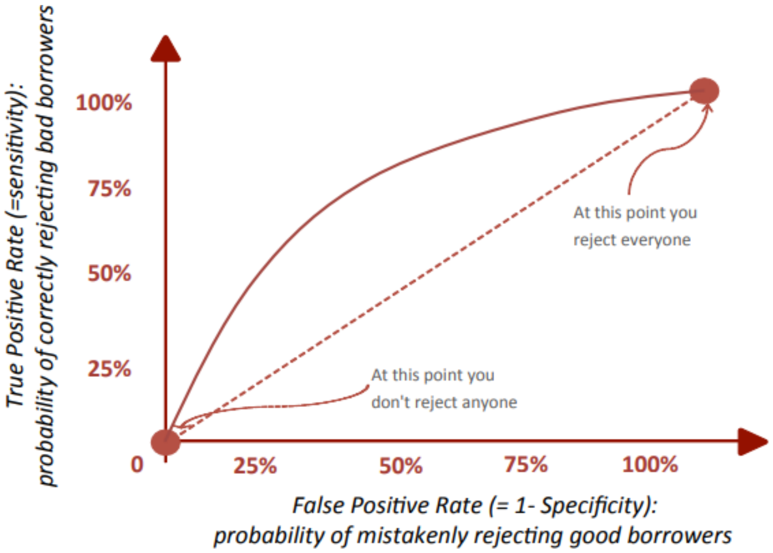
For the analyzation, the logarithm is used for the variables *applied amount* and *income*. The use of log is to prevent the regression from skewness towards outlining values[5]. Another important reason for implementing a logarithm is to help rescale the data so that the variance is more constant. The rescaling of the data is to overcome the heteroscedasticity problem and make the model more homoscedastic. Heteroscedasticity occurs when the variance of the dependent variable is dependent on the variance of the independent variable.

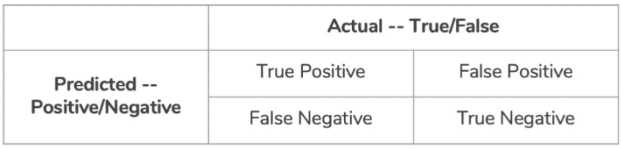
[4] https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/fixed-effects-regression

[5] https://www.forbes.com/sites/naomirobbins/2012/01/19/when-should-i-use-logarithmic-scales-in-my-charts-and-graphs/#:~:text=There%20are%20two%20main%20reasons,percent%20change%20or%20multiplicative%20factors.

## 4.2 ROC curve

LendingClub's model is tested at the percentage of accuracy in the classification. A standard measure of the classification accuracy is the Receiver Operating Characteristic (ROC). The Receiver Operating Characteristics (ROC) curve is a popular technique to validate (internal) credit rating models and investigate the area under the curve to qualify the model (Agarwal & Taffler, 2008). Sobehart and Keenan (2001) explain how to implement a ROC curve to verify the model and show that it has an unbiased estimator. ROC indicates the partially amount of predicting default on loans that are indeed classified as default (sensitivity) against the amount of non-default loans falsely classified as default (1-specificity). The output of the ROC lies between 0.5 and 1, whereas 0.5 indicates a "coin flip" and 1 means a perfect forecast without errors. The ROC-curve is determined for the *DTI* ratio and the *Rating* of LendingClub. An example is shown in Figure 3.





| | Actual -- True/False | |
|---|---|---|
| Predicted -- Positive/Negative | True Positive | False Positive |
| | False Negative | True Negative |

*Figure 3: a ROC curve. Adopted from Iyer et al. (2009).*

The further process is as follows. A ROC curve is defined by the risk of default from highest to lowest risk for all the loans. For every loan, a default rate is calculated on the independent variables regarding the regression in paragraph 4.1. Afterwards, the total number of defaults is divided by all defaults within the dataset and plotted against all loans.

Consider the table above, where the value to determine is default. In a perfect world, only the "True Positive" (T.P.) and "True Negative" (T.N.) will occur since the model is correctly predicting. But, in the model used in this thesis, this will not be the case. The performance of the regression model in this research will be analyzed due to the sensitivity ratio, namely the percentage of correctly forecasted "T.P." and "T.N.".

ROC curves are evaluated based on the area under the curve (AUC), which have a range between 0.5 and 1. According to Iyer et al. (2009), a general rule is that an AUC of 0.6 or greater is believed that is suitable in scarce information models, whereas an AUC of 0.7 or more is the aim of enricher models.

## 4.3 Logistic regression

Logistic or logit regression measures the relationship between the categorical or binary dependent variable and one or multiple independent variables by calculating probabilities. The purpose of the logistic regression is to analyze whether the variable(s) in the model has a significant explanatory effect. Logit regression is similar to Ordinary Least Squares but has the advantage of dealing with a binary dependent variable. Logit model uses a model non-linear in parameters to estimate the probability of an event to occur. The estimation method used is the Maximum Likelihood Estimation (MLE).

### 4.3.1 Marginal effects

Marginal effects of a variable are the proportion of x at which a dependent variable change at a given point in the regression, concerning a variable and holding other factors fixed. This outcome is useful because it provides insight into the influence of a variable on the prediction of the dependent variable *Default*. Marginal effects are calculated due to partial derivatives. The slope of a multi-dimensional regression regarding one variable is the marginal effect. Within OLS, there are three options for selecting marginal effects, namely: Marginal effects at representative values (MERs), Marginal effects at means (MEMs), and Average marginal effects (AMEs).

MERs determine the marginal impact of each variable at a group of independent variables. MEMs are focusing on each variable at the mean of the covariates. AMEs compute marginal effects based on every observation and average through the effect estimates.

This dissertation concentrates on AMEs since AMEs estimate one quantity summary that displays the full sample of the variable rather than a random prognosticating, as happens with MEMs. AMEs have a preference over MERs because AMEs offers an outcome that corresponds to one variable, whereas MERs looks towards typical values. AMEs are better to use since this model uses categorical dummies and time fixed effects. AME computes the probability to default for every category and takes the difference of this estimated probability to calculate the AME.

# 5. Results

This section provides the results of analyzing the default risk of the dataset of LendingClub between 2009 and 2019. First, subsection 5.1 shows the correlation matrix of all the incorporated variables in the model. Secondly, the linear regression of the model is discussed in subsection 4.2. Thirdly, subsection 4.3 provides the prediction of default for the dataset of LendingClub. Finally, subsection 4.4 analyzes the ROC curve of the model.

## 5.1 Correlation matrix and frequency table

Table 3 displays an overview of the correlation between the two variables of the model. Almost all variables are statistically significant with each other at 1%, except for the relation between *Verified* and *House*. As stated in Table 3, *Loan amount, loan term, interest rate, installment, DTI, and Verified* have a positive correlation with *Default*. On the other hand, *log (Income), FICO range,* and *House* have a negative correlation with the dependent variable *Default*.

Subsequently, Table 3 is included to test for multicollinearity, which could happen when regressions are testing the hypothesis. Multicollinearity is the problem of having two or more explanatory variables that are highly (but not perfectly) correlated. Multicollinearity does not generate biased OLS estimates, but it increases the standard errors or the correlated variables. The variables *log (Applied Amount), Loan term, Interest rate* and *log (Income)* are around 40 percent correlated with each other, but this is not enough to be at risk for multicollinearity. All other variables are not correlated with each other and are easily addable towards the regression without any problem of multicollinearity.

In Table 4, the frequencies of the number of loans within several grades are displayed and provides more insight into the difference within LendingClub's *Grades* and the default risk. The lower the provided *Grade* by LendingClub, the higher the percentage of *Default* within the loans (shown in the second row of each *Grade* level). At *Grade* level A, only 5.73% of the loans happen to be charged off, whereas, at *Grade* level F and G, 41.27% and 46.93% of the loans are defaulted, respectively. Most loans are falling in the group of *Grade* B or C, which consists of almost 60% of the loans.

## 5.2 Logit model

Table 5 shows the logit model for the forecasting of default for all variables without the inclusion of dummies. All variables are statistically significant, which indicates that the variables have a substantial influence on the dependent variable (*Default*). *Applied Amount*, *Interest, DTI* and *FICO score* all have a positive effect on the dependent variable, so the higher significant the input of these, the more probability of default. *Log (Income)*, *House*, *Verified, Loan term* all have a negative effect on the dependent variable.

Additionally, a useful way to interpret the table is the likelihood value which is shown in the last row of the table. The likelihood measures how likely it is that the giving regression is providing a dataset as the analyzing dataset. Thus, the higher the likelihood value, the better the model will fit. This measurement is called "Maximum Likelihood Estimation" (MLE).

Table 6 shows the results of the complete logistic regression made to determine the estimators of *Default*. In Table 6, two dummies (*Employment* and *Grade)* are included. Within the regression, *FICO score* is the variable with the most impact and has a negative correlation with the risk of *Default*, which means that the higher the credit score, the lower the risk of default. This is in line with a study of Emekter, Tu, Jirasakuldech, and Lu (2015), which shows that credit grade, *DTI* and *FICO score* have an impact on the risk of default. Low credit grade and longer maturity are correlated with a higher default rate. *DTI* has a profoundly positive effect on the risk of *Default*; this might be because it is more difficult for borrowers to be financially stable when having more debt related to income. *FICO score* has a significant adverse effect on the *Default* risk, so an increase in *FICO score* provides a higher probability of repaying the loan. *Verified* has a slightly less negative impact on the prediction of *Default.* Thus borrowers who prove their income source are less expected to default on their loan. Referring to Kumar (2007), income verification is correlated with lower default risk, and this is validated by the regression conducted. The (log) amount of the loan (*log (Amount Applied*)) has a positive correlation with the dependent variable. Kumar (2007) states that more significant loans are easier being defaulted because these loans are harder to be repaid.

Another interesting variable is the *Interest rate,* which has a relatively high positive correlation with the risk of *Default.* An increase in interest rate indicates a higher risk, and, therefore, a greater possibility of default, which is related to the paper by Dorfleitner et al. (2017). *Loan term* also has a positive relationship with the probability on default, illustrating that longer maturity increases the risk of default. Besides, *Employment* has no statistically significant effect, so there can not be concluded anything on the relationship between working experience and *Default*. Except for more than ten years of experience, with a significance level of 10%, that has a positive relation towards *Default*. Additionally, *Housing* status relates negatively to the status of *Default*. This outcome indicates that borrowers with a mortgage have a lower probability of default on their loan.

One of the most important variables, namely *Grade*, exhibit a positive relationship regarding *Default*. Borrowers with a higher grade have a lower possibility of default. *Grade* A has a negative slope, determining that creditors with an A grade are more reliable on repaying their loan. Borrowers with a B grade are scoring better than with a C grade, respectively 0.0222 against 0.0628.

The numbers are designed to be interpreted as percentage points. So, borrowers with C grade have 4.04 percentage point (0.0628 – 0.0222) more probability of not repaying their loan. Overall, LendingClub's *Grades* are, indeed, correct forecasters in terms of *Default*, which is corresponding with the paper of Jagtiani & Lemieux (2019). Nonetheless, the relationship between *Grades* and the *FICO score*s is not proven in this regression.

Furthermore, in Table 7, an interaction term between *DTI* and low credit borrowers (*Grade* D till G) is conducted. In Table 7, *DTI* also has a positive relation towards Default as the probability of Default increases when *DTI* increases. This positive relation makes economic sense since higher debt-to-income leads to a higher risk of default. A *DTI* between 30 till 40 indicates that the chance of Default is 7.31 percentage point higher than a *DTI* ratio from zero to ten.

Specification (3), (4), and (5) compare the complete sample with only default or repaid loans with the additional interaction term. Borrowers with a low rating have 14.8 percentage points more likely to *Default* than debtors with a high rating. Borrowers with a low rating and a *DTI* between zero to ten have an additional 0.74 percentage point higher chance of *Default*. In comparison, low rating borrowers with a *DTI* between 25 to 30 have 3.33 percentage point higher probability of *Default*. The difference between default and repaid loans is that default loans have overall a more increasing coefficient for all variables than repaid loans, and, on average, a higher standard error which indicates less precise estimators.

Specification (5) shows the Average Marginal Effects (AME) of the relationship between *Default*, *DTI*, and a low credit borrower. According to Table 7, the interaction term of borrowers with a *DTI* of 40 or higher and a low rating have 33.87 percentage point more probability of *Default* than borrowers with a *DTI* between zero to ten. To compare, a *DTI* score from 10 to 15 is 8.91 percentage point higher than the base group. These results are as expected and are in line with the previous Tables 5 and 6. All variables are statistically significant.

## 5.3 Interpret the ROC and predict default

Receiver Operating Characteristics (ROC) curve is implemented for credit rating models and investigations of the area under the curve (AUC) to qualify the model. Graph 1 displays the ROC curves for the relationship between *Default* and different variables. The AUC related to *Interest* is 0.6776, and the AUC of *FICO score* is 0.5911. This result indicates that interest rate is a better default forecaster than *FICO score*. As mentioned in section 4.2, an AUC of 0.6 or above is a general measure in less obvious expected environments, for instance, screening of small borrowers. The AUC for the grade, income, employment and all variables together are 0.6756, 0.5469, 0.5145, and 0.7031, respectively.

Graph 1 illustrates that employment and income are less helpful in terms of *Default* prediction than grade and interest rate. Forecast of *Default* can be concluded with an accuracy ratio of 70.31% within in this dissertation's model.

Complete uninformative information has an AUC of 0.5, and even 0.01 expansion in AUC is a relevant and sufficient gain in the credit- or lending industry (Iyer et al., 2019). Iyer et al. (2019) mention that lenders can forecast default with 45% more accuracy than by only using the credit score, an increase that is significant at the 1% significance.

# 6. Discussion

This section gives a general discussion of this research, and recommendations for further research are provided. First, a little background is stated to introduce the study in subsection 6.1. Secondly, the results and conclusions of the different hypothesis are formulated in subsection 6.2. Finally, subsection 6.3 mentions limitations on this research project and suggests several directions for further research.

## 6.1 Background and research question

This thesis investigated the financial characteristics of P2P lending and the question of whether default can be predicted. Peer-to-peer (P2P) lending describes the loan origination between private individuals (peers) on online platforms where financial institutions operate only as intermediates required by law (Bachmann et al., 2011). The data is from LendingClub, a United States P2P platform, and has a timeframe from 2009 till 2016.

Recently, P2P lending is developed as a new form of credit within FinTech and competes with traditional banks. But, the borrowers of P2P platforms consist of riskier borrowers which result in higher default rates (de Roure, Pelizzon, & Thakor, 2018). Several studies mention that characteristics as interest rate, debt-to-income, and credit rating influence the default risk (Dorfleitner et al. (2017); Klafft (2008); Serrano-Cinca et al. (2015)), but none of these researches examines the forecasting of default.

Prediction of default, however, is essential for lenders and this new technology which uses algorithms to evaluate the trustworthiness of applicants and prevent information asymmetry. Empirical results are based on logistic regressions, and, besides, a ROC curve is investigated to determine the effect of different variables and examines the accuracy of this dissertation's model for default loans.

## 6.2 Results and conclusions

To answer the general research question *What are the financial determinants of default within P2P lending, and can we model the default rate?* three hypotheses are explored to interpret the research question clearer. The following results were found.

**Hypothesis (1):** *Which financial information have an impact on the default risk?*

After running different logistic regressions, *Applied Amount*, *Interest, DTI* and *FICO score* all have a positive effect on the dependent variable as displayed in Table 5 and 6, which is in line with multiple papers (Dorfleitner et al., 2017; Klafft, 2008; Serrano-Cinca et al., 2015). Higher monthly payments require more income and responsibility for the borrower and cause a higher risk of default. Results show that lenders are providing loans with higher interest rates when *DTI* and *Applied Amount* are higher. *DTI* indicates the capability of a borrower of paying back their debt and shows the risk tolerance of these loan applicants. *Default* and *DTI* have a positive relationship because a higher *DTI* indicates less capital to repay their debt and thus more risk of defaulting their requested loan. Lenders rely on several variables, whereby a higher *FICO score* incorporates more trust and financial stability, and, therefore, a lower risk of default. *FICO score* is computed by an independent company in the United States, which suggest that lenders could rely on this score to determine the trustworthiness of a borrower.

On the other hand, *Log (Income)*, *House*, *Verified,* and *Loan term* all harm the dependent variable, which also corresponds to various studies (Chen et al., 2019; Jin and Zhu, 2015). Empirical regressions show that certain variables can assess the creditworthiness of borrowers. If the income of a borrower is lower, less financial sources are available for the repayment of their applied loan. Longer maturity corresponds to more time for financial mistakes that can fail payments. Besides, verification of income sources is essential in terms of trustability and confidence about the borrower's financial position. Debtors that are not willing to verify their financial resources might be aware of their inability to repay and, therefore, are profiting by information asymmetry. Homeowners are less risky than renters because house owners have an asset they can rely on and have a mortgage to demonstrate that these borrowers are more trustworthy.

Working experience has no statistically significant effect, so there is no statistical effect to be concluded. LendingClub's *Grade* has an increasingly positive influence on the default risk, as shown in Table 6. Higher provided grades by LendingClub decreases the risk of default and, therefore, positively effects the dependent variable. Results show that LendingClub's *Grade* is relatively accurate in terms of predicting failure of payment, and is significant in terms of the loan- and borrower characteristics.

**Hypothesis (2):** *Can the default rate be forecasted based on these financial determinants?*

Forecasting the default risk is constructed by the ROC curve and the AUC when the independent variable has a binary outcome. The model stated in section 4 is examined to answer this hypothesis. The AUC regarding interest rate is 0.6776, and the AUC of *FICO score* is 0.5911. As mentioned in section 4.2, an AUC of 0.6 or above is a general measure in the financial world, especially for small borrower prediction (Iyer et al., 2009). The AUC for the grade, income, employment and all variables together are 0.6756, 0.5469, 0.5145, and 0.7031, respectively. *Grade*, which is provided by LendingClub, is the most accurate variable referring to the default calculation, whereas employment length have a insignificant effect. Denoting the variable *Income*, an AUC score of 54.69% is not appropriate enough to conclude the impact of *Income* on the forecasting of *Default.*

*Income* is provided by the borrower, which may result in an uncertain number and, thus, is not reliable in the prediction analysis.

Default risk is still a challenging outcome to prognosticate. According to Graph 1, *Default* can be forecasted with an accuracy ratio of 70.31% for all the variables in the model. The average default rate of all loans is 17.33%, which explains that loans from LendingClub are at a high-risk level. Variables like *DTI*, *FICO score*, and *Interest* have a wide range of possibilities and high standard errors, and this makes it hard to foresee correctly. Borrowers at LendingClub are more risk seekers and are financially less stable, resulting in less precise prediction and more charged off loans.

Overall, the financial characteristics determined by this thesis's model are an adequate indicator to forecast default. The accuracy ratio of 70.31% is, in terms of the small borrower forecast, above the general threshold of 60%. All financial determinants are statistically significant with 1% and have a specific effect on the dependent variable.

**Hypothesis (3):** *Can a computer-based algorithm outperform the system of LendingClub in terms of default prediction?*

The difference between this dissertation's model and the system of LendingClub is the inclusion of various independent variables and fixed effects. Within LendingClub, there is a list of almost 150 variables which are included in the model, whereby the importance of different variables is not mentioned[6]. In this thesis, the model is exhibit under self-selected variables which are all statistically significant with 1%.

Overall, this computer-built algorithm has a sufficient better forecast accuracy (70.33%) than the overall LendingClub accuracy (63.56%). LendingClub incorporates more variables, which includes that precision can be less accurate. The model of this dissertation had different dummies which are dependent on LendingClub's self-provided *Grade* and borrower's employment years.

## 6.3 Limitations and future research

In closing, there were some limitations to this study. First of all, the full dataset was including "Current" loans which are still in payment and are neglected out of this study. These loans are appealing because FinTech is changing rapidly over time and thus a comparison of the years 2017 and further would be interesting.

Secondly, this thesis could use alternative benchmarks, like conducting a ROC curve as an econometrician and conduct with all the available variables, and all the available textual information as loan purpose. Future research could, for this reason, incorporate these benchmarks to have more comparison between variables.

[6] https://www.lendingclub.com/statistics/additional-statistics?

Thirdly, future research could examine the initial moment of *Default* and investigate if there is a relationship between individual events. An event study could be conducted to address this in future research. Different quarters of years would provide other coefficients and could be more accurate in terms of prediction.

Lastly, a comparison between different P2P platforms could be investigated, as LendingClub is independent and is located in the United States. Future research could examine P2P platforms from other economies to provide deeper insights into financial determinants or forecasting results within FinTech.

# 7. Conclusion

Since the banking crisis of 2008, the barriers of credit access have increased in the United States and Europe (Calabrese, Osmetti, & Zanin, 2019), and peer-to-peer (P2P) lending is rising more and more within the FinTech (financial technology) world. Despite the studies on P2P lending and new FinTech techniques, there is little research examined on the prediction of default and the risk within P2P platforms. The dissertation examines data from LendingClub, a United States founded P2P platform, between 2009 and 2016.

This research project investigates the different financial characteristics determined to analyze the default risk and examines if the default rate can be predicted. One of the key findings is that financial traits which have a positive relationship with the default rate are debt-to-income (*DTI*), the maturity of the loan, and the interest rate. Secondly, financial determinants with a negative effect are applied amount, income, FICO score, the ownership of a house, and the verification of the income (source). Thirdly, variables with more accuracy in terms of prediction are income, interest rate, and DTI. Lastly, the model established in this thesis is better in predicting default than the model of LendingClub with 70.33% against 63.65% accuracy.

To conclude, default risk can be predicted and analyzed with various financial determinant. Different variables and time frames are influencing the default risk, and, thus, future research should investigate deeper on particular variables and more timeframes when borrowers tend to default.

# References

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8), 1541-1551.

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., & Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, 16(2).

Balyuk, T. (2016). Financial Innovation and Borrowers. *Rotman School of Management*.

Bricker, J., Dettling, L. J., Henriques, A., Hsu, J. W., Jacobs, L., Moore, K. B., . . . Windle, R. A. (2017). Changes in U.S. Family Finances from 2013 to 2016: Evidence from the Survey of Consumer Finances. *Federal Reserve Bulletin*, 103(3):1–42.

Buchak, G., Matvos, G., Piskorski, T., & Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics*, 130(3), 453-483.

Calabrese, R., Osmetti, S. A., & Zanin, L. (2019). A joint scoring model for peer-to-peer and traditional lending: a bivariate model with copula dependence. *Journal of the Royal Statistical Society*, 182(4), 1163-1188.

Chen, C. W., Dong, M. C., Liu, N., & Sriboonchitta, S. (2019). Inferences of default risk and borrower characteristics on P2P lending. *The North American Journal of Economics and Finance*, 50, 101013.

Chishti, S. (2016). How peer to peer lending and crowdfunding drive the FinTech revolution in the U.K. *Baking Beyond Banks and Money*, 55-68.

Croux, C., Jagtiani, J., Korivi, T., & Vulanovic, M. (2020). Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform. *Journal of Economic Behavior & Organization*, 173, 270-296.

de Roure, C., Pelizzon, L., & Thakor, A. (2018). P2P lenders versus banks: Cream. *SAFE Working Paper No. 206*.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2017). Description-text related soft information in peer-to-peer lending. *Journal of Banking & Finance, 64*, 169-187.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, 25(8), 2455-2484.

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1), 54-70.

European Commission. (2018). *FinTech Action plan: For a more competitive and innovative European financial sector.* European Commission.

Feng, Y., Fan, X., & Yoon, Y. (2015). Lenders and borrowers' strategies in the online peer-to-peer lending market: an empirical analysis. *Journal of Electronic Commerce Research*, 16(3), 242.

Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2019). The role of technology in mortgage lending. *Review of Financial Studies*, 32(5), 1854-1899.

Gao, Q., & Lin, M. (2013). Linguistic Features and Peer-to-Peer Loan Quality: A Machine. *SSRN Electronic Journal*.

Guild, J. (2017). Fintech and the Future of Finance. *Asian Journal of Public Affairs*, 52-65.

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? *Management Science*.

Jagtiani, J., & Lemieux, C. (2018). Do fintech lenders penetrate areas that are underserved by traditional banks? *Journal of Economics and Business*, 100, 43-54.

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. *Financial Management*, 48(4), 1009-1029.

Jagtiani, J., Lambie-Hanson, L., & Lambie-Hanson, T. (2019). Fintech lending and mortgage credit access. *Philadelphia Fed Working Paper*.

Jensen, M. C., & Meckling, W. H. (1979). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Economics Social Institutions*, 163-231.

Jiang, E., Matvos, G., Piskorski, T., & Seru, A. (2020). Banking without deposits: Evidence from shadow bank call reports. *National Bureau of Economic Research.*

Jin, Y., & Zhu, Y. (2015). A data-driven approach to predict default risk of the loan for online peer-to-peer (P2P) lending. *Fifth International Conference on Communication Systems and Network Technologies*, (pp. 609-613). IEEE.

Käfer, B. (2018). Peer-to-Peer Lending–A (Financial Stability) Risk Perspective. *Review of Economics*, 69(1), 1-25.

Klafft, M. (2008). Peer to peer lending: auctioning microcredits over the internet. *International Conference on Information Systems, Technology and Management*.

Kotter, J., & Lel, U. (2011). Friends or Foes? Target selection decisions of sovereign wealth funds and their consequences. *Journal of Financial Economics, 101*(2), 360-381.

Kumar, S. (2007). Bank of one: Empirical analysis of peer-to-peer financial marketplaces. *AMCIS 2007 Proceedings*, 305.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1), 19-37.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1), 17-35.

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. *Applied Economics*, 49(35), 3538-3545.

Michels, J. (2012). Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *The Accounting Review*, 87(4), 1385-1413.

Milne, A., & Parboteeah, P. (2016). The business models and economics of peer-to-peer lending. *European Credit Research Institute*.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLoS One*, 10(10).

Sobehart, J., & Keenan, S. (2001). Measuring default accurately. *Risk*, 14(3), 31-33.

Tang, H. (2019). Peer-to-Peer Lenders versus Banks: Substitues or Complements? *Review of Financial Studies*, 1900-1938.

Yum, H., Lee, B., & 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), 469-483.

# Tables and Graphs

## Table 1 Variable definitions

This table provides an overview of all the variables used in this thesis. The variables are split in loan results, loan characteristics and borrower characteristics.

| **Variable** | **Label** | **Definition** |
|---|---|---|
| *Loan results* | | |
| default | Default | 1 if the loan has been defaulted and 0 otherwise |
| *Loan characteristics* | | |
| loan_amnt | Applied amount | Loan amount applied by the borrower |
| term | Loan term | Loan duration agreed upon in loan contract |
| int_rate | Interest rate | The interest rate of the loan |
| *Borrower characteristics* | | |
| monthly_inc | Income | Total monthly income of a borrower |
| verification_status | Verified | The method used in the loan application process to verify income 1 if income (source) is verified and 0 otherwise |
| emp_length | Employment length | Employment length in years. |
| grade | Grade | Rating calculated by LendingClub's rating model ranging from 1 (= A = very low risk) to 7 (= G = very high risk). |
| home_owner | House | 1 if the borrower is a homeowner or has a mortgage and 0 otherwise |
| addr_state | State | Residence state of the borrower |
| dti | Debt-to-income | A ratio calculated by using monthly debt payments (excluding mortgage and LendingClub’s loan) divided by the borrower’s reported monthly income. |
| fico_score | *FICO score* | The borrower's *FICO score* |
| installment | Installment | The monthly payment owed by the borrower if the loan originates. |

## Table 2.1 Descriptive statistics

The table shows the summary statistics of all provided loans within the platform LendingClub between 2009 and 2016. The first column is the total observations; the second column indicates the average of all loans, the third column shows the median, the fourth column the standard deviation (S.D.), the fifth column the minimum value, and the last column the maximum value of the variable.

| | (1) | (2) | (3) | (3) | (4) | (5) |
|---|---|---|---|---|---|---|
| Variables | N | mean | median | S.D. | min | max |
| *Loan results* | | | | | | |
| default | 725,641 | 0.1733 | 0 | 0.3786 | 0 | 1 |
| | | | | | | |
| *Loan characteristics* | | | | | | |
| int_rate | 725,641 | 0.1323 | 0.1284 | 0.0455 | 0.0532 | 0.310 |
| loan_amnt | 725,641 | 14,333 | 12,000 | 8,555 | 1,000 | 40,000 |
| term | 725,641 | 42.01 | 36 | 10.40 | 36 | 60 |
| | | | | | | |
| *Borrower characteristics* | | | | | | |
| installment | 725,641 | 433.0 | 374.33 | 254.5 | 4.930 | 1,585 |
| annual_inc | 725,558 | 75,494 | 65,000 | 68,983 | 0 | 9,573,072 |
| dti | 721,164 | 18.06 | 17.54 | 9.298 | 0 | 99 |
| fico_score | 721,200 | 696.0 | 690 | 31.16 | 660 | 845 |
| verification_factor | 725,558 | 2.043 | 2 | 0.794 | 1 | 3 |

## Table 2.2 Descriptive statistics for default or repaid

The table shows the summary statistics of all provided loans within the platform LendingClub between 2009 and 2016. Default (default 1) means that the loan is default and repaid (default 0) includes all the loans which are fully paid. All variables are defined in Table 1.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | default 0 | | | | | | default 1 | | | | | |
| VARIABLES | N | mean | median | sd | min | max | N | mean | median | sd | min | max |
| loan_amnt | 594,256 | 14,084 | 12,000 | 8,517 | 1,000 | 40,000 | 131,385 | 15,462 | 14,325 | 8,635 | 1,000 | 40,000 |
| term | 594,256 | 41.14 | 36 | 9.848 | 36 | 60 | 131,385 | 45.91 | 36 | 11.82 | 36 | 60 |
| int_rate | 594,256 | 0.127 | 0.1229 | 0.0436 | 0.0532 | 0.310 | 131,385 | 0.155 | 14.99 | 0.0464 | 0.0532 | 0.310 |
| installment | 594,256 | 427.8 | 368.65 | 254.7 | 4.930 | 1,585 | 131,385 | 456.2 | 399.72 | 252.4 | 21.62 | 1,585 |
| annual_inc | 594,256 | 76,785 | 65,000 | 70,194 | 0 | 9,225,000 | 131,302 | 69,653 | 60,000 | 62,883 | 0 | 9,573,072 |
| dti | 590,538 | 17.67 | 17.1 | 8.976 | 0 | 999 | 130,626 | 19.85 | 19.56 | 10.44 | 0 | 999 |
| fico_score | 590,567 | 699.8 | 692 | 32.02 | 662 | 848 | 130,633 | 689.8 | 682 | 25.38 | 662 | 848 |
| homeowner_factor | 594,256 | 2.289 | 2 | 0.642 | 1 | 6 | 131,302 | 2.351 | 2 | 0.666 | 1 | 6 |
| verification_factor | 594,256 | 2.014 | 2 | 0.798 | 1 | 3 | 131,302 | 2.173 | 2 | 0.765 | 1 | 3 |
| emp_years | 594,256 | 4.986 | 4 | 3.480 | 1 | 12 | 131,302 | 5.198 | 4 | 3.610 | 1 | 12 |

## Table 3 Correlation matrix

This table shows the correlation between the used variables in the dataset. Standard errors are clustered at the borrower level and are displayed with *, **, and *** to specify statistical significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | default | loan_amnt_log | term | int_rate | installment_low | income_log | dti | fico_score | lhouse | dummy_verified |
| default | 1 | | | | | | | | | |
| loan_amnt_log | 0.0657$^{***}$ | 1 | | | | | | | | |
| Term | 0.1776$^{***}$ | 0.3902$^{***}$ | 1 | | | | | | | |
| int_rate | 0.2440$^{***}$ | 0.1120$^{***}$ | 0.4189$^{***}$ | 1 | | | | | | |
| installment_low | 0.0428$^{***}$ | 0.2833$^{***}$ | 0.1432$^{***}$ | 0.1493$^{***}$ | 1 | | | | | |
| income_log | -0.0615$^{***}$ | 0.3713$^{***}$ | 0.1252$^{***}$ | -0.1152$^{***}$ | 0.3887$^{***}$ | 1 | | | | |
| dti | 0.0904$^{***}$ | 0.0411$^{***}$ | 0.0714$^{***}$ | 0.1574$^{***}$ | 0.0354$^{***}$ | -0.2090$^{***}$ | 1 | | | |
| fico_score | -0.1230$^{***}$ | 0.0872$^{***}$ | -0.0047$^{***}$ | -0.426$^{***}$ | 0.0545$^{***}$ | 0.1062$^{***}$ | -0.0849$^{***}$ | 1 | | |
| lhouse | -0.0551$^{***}$ | 0.1769$^{***}$ | 0.1087$^{***}$ | -0.0736$^{***}$ | 0.1534$^{***}$ | 0.2697$^{***}$ | -0.0006 | 0.1046$^{***}$ | 1 | |
| dummy_verified | 0.0759$^{***}$ | 0.1769$^{***}$ | 0.1456$^{***}$ | 0.2218$^{***}$ | 0.2050$^{***}$ | 0.0717$^{***}$ | 0.0515$^{***}$ | -0.1488$^{***}$ | -0.0011 | 1 |

## Table 4 Frequency table *Grade* on *Default*

Table 4 illustrates the frequencies and percentages of different *Grade*s and the risk of *Default.*

| | Default | | |
|---|---:|---:|---:|
| Grade | 0 | 1 | Total |
| A | 117,926 | 7,163 | 125,089 |
| | 94.27 | 5.73 | 100.00 |
| | 19.84 | 5.45 | 17.24 |
| B | 193,698 | 26,775 | 220,473 |
| | 87.86 | 12.14 | 100.00 |
| | 32.60 | 20.38 | 30.38 |
| C | 160,736 | 41,158 | 201,894 |
| | 79.61 | 20.39 | 100.00 |
| | 27.05 | 31.33 | 27.82 |
| D | 77,498 | 29,540 | 107,038 |
| | 72.40 | 27.60 | 100.00 |
| | 13.04 | 22.48 | 14.75 |
| E | 31,951 | 17,589 | 49,540 |
| | 64.50 | 35.50 | 100.00 |
| | 5.38 | 13.39 | 6.83 |
| F | 10,164 | 7,141 | 17,305 |
| | 58.73 | 41.27 | 100.00 |
| | 1.71 | 5.44 | 2.38 |
| G | 2,283 | 2,019 | 4,302 |
| | 53.07 | 46.93 | 100.00 |
| | 0.38 | 1.54 | 0.59 |
| Total | 594,256 | 131,385 | 725,641 |
| | 81.89 | 18.11 | 100.00 |
| | 100.00 | 100.00 | 100.00 |

First row has *frequencies*; second row has *row percentages* and third row has *column percentages*

## Table 5 Regression output 1

This table demonstrates the results of the logistic regression output, which analyzes the relationship between *Default* and the independent variables. All loans from LendingClub between 2009 and 2016 are included in the sample. Fixed effects are added in the regression as displayed. Specification 1 shows the regression coefficients of the baseline model. In specification 2, the coefficients are displayed where only defaulted loans are included. Specification 3 covers the coefficients where only repaid loans are analyzed. Specification 4 indicates the AMEs for every variable. All variables are defined in Table 1. Standard errors are clustered at the borrower level and are noted in parentheses, and *, **, and *** show statistical significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| VARIABLES | Baseline | Only default loans | Only repaid loans | Average Marginal Effects |
| Log (Applied amount) | -0.0124*** | -0.0314*** | -0.0087*** | -0.0381*** |
| | (0.0005) | (0.0008) | (0.0003) | (0.0215) |
| Loan term | 0.0734*** | 0.1194*** | 0.0612*** | 0.0579*** |
| | (0.0008) | (0.0021) | (0.0009) | (0.0048) |
| Interest rate | 0.1819*** | 0.2312*** | 0.1691*** | 0.1419*** |
| | (0.0151) | (0.0121) | (0.0081) | (0.0159) |
| Log (Income) | -0.1352*** | -0.1143*** | -0.1471*** | -0.3433*** |
| | (0.0251) | (0.0311) | (0.0402) | (0.0651) |
| Verified | -0.1832*** | -0.2274*** | -0.1722*** | -0.2132*** |
| | (0.0055) | (0.0084) | (0.0062) | (0.0125) |
| *FICO score* | -0.2145*** | -0.1823*** | -0.2742*** | -0.2632*** |
| | (0.0142) | (0.0294) | (0.0121) | (0.0421) |
| House | -0.2056*** | -0.2123*** | -0.2297*** | -0.3656*** |
| | (0.0336) | (0.0527) | (0.0429) | (0.0522) |
| *DTI* | 0.2901*** | 0.3412*** | 0.2437*** | 0.3287*** |
| | (0.0658) | (0.0894) | (0.0512) | (0.0958) |
| Constant | 0.2021*** | 0.3531*** | 0.0921*** | |
| | (0.0647) | (0.1089) | (0.0664) | |
| Year F.E. | Yes | Yes | Yes | |
| Observations | 725,641 | 131,385 | 594,256 | |
| $R^2$ | 0.4124 | 0.5126 | 0.5433 | |
| Log likelihood | -312323.44 | -341186.95 | -352341.95 | |

## Table 6 Regression output 2

This table displays the results of the complete regression, which analyzes the relationship between *Default* and the independent variables of the regression. This model includes a dummy for *Employment* which contains the working years of the applicant and a dummy for *Grade,* which can take values between A (high) to G (low). Fixed effects are included in the analyzation as displayed. Specification 1 shows the regression coefficients of the baseline model. In specification 2, the coefficients are shown where only defaulted loans are included. Specification 3 displays the coefficients where only repaid loans are examined.

Additionally, specification 4 indicates the AMEs for every variable. All variables are defined in Table 1. Standard errors are clustered at the borrower level and are displayed in parentheses, and *, **, and *** specify statistical significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| VARIABLES | Baseline | Only default loans | Only repaid loans | Average Marginal Effects |
| Log (Applied Amount) | -0.0205*** | -0.0468*** | -0.0104*** | -0.0247*** |
| | (0.0012) | (0.0011) | (0.0005) | (0.0013) |
| Loan term | 0.0732*** | 0.1242*** | 0.0413*** | 0.01845*** |
| | (0.0000) | (0.0037) | (0.0087) | (0.0007) |
| Interest rate | 0.2541*** | 0.3180*** | 0.2411*** | 0.2751*** |
| | (0.0312) | (0.0045) | (0.0006) | (0.0121) |
| Log (Income) | -0.0861*** | -0.1250*** | -0.1321*** | -0.1261*** |
| | (0.0010) | (0.0053) | (0.0052) | (0.0156) |
| Verified | -0.1142*** | -0.1923*** | -0.0812*** | -0.2155*** |
| | (0.0210) | (0.0208) | (0.0113) | (0.0240) |
| *FICO score* | -0.2369*** | -0.2752*** | -0.1822*** | -0.1434*** |
| | (0.0142) | (0.0233) | (0.0121) | (0.0130) |
| House | -0.2512*** | -0.2832*** | -0.2123*** | -0.2675*** |
| | (0.0019) | (0.0312) | (0.0153) | (0.0465) |
| *DTI* | 0.1178*** | 0.1891*** | 0.1233 *** | 0.1253*** |
| | (0.0101) | (0.0165) | (0.0024) | (0.0141) |
| *Employment dummy* | | | | |
| < one year | -0.0466 | -0.0511 | -0.0351 | -0.0147 |
| | (0.1134) | (0.1244) | (0.1133) | (0.0309) |
| < three years | 0.0893 | 0.0983 | 0.0765 | 0.01694 |
| | (0.1323) | (0.1123) | (0.1344) | (0.0348) |
| < five years | 0.1132 | 0.0775 | 0.0988 | 0.0324 |
| | (0.2101) | (0.1423) | (0.1358) | (0.0147) |
| < seven years | 0.0803 | 0.1153 | 0.0942 | 0.03548 |
| | (0.1626) | (0.1342) | (0.1456) | (0.0836) |
| < ten years | 0.1332 | 0.2132 | 0.1457 | -0.0934 |
| | (0.1516) | (0.1821) | (0.1685) | (0.0648) |
| > ten years | 0.1511* | 0.1732* | 0.1487* | 0.0845** |
| | (0.0897) | (0.0657) | (0.0509) | (0.0312) |
| *Rating dummy* | | | | |
| Rating A | -0.0345*** | -0.0921*** | -0.0511*** | -0.0884*** |
| | (0.0087) | (0.0169) | (0.0078) | (0.0041) |
| Rating B | 0.0222*** | -0.0215 | 0.0017 | 0.0341*** |
| | (0.0014) | (0.0921) | (0.0045) | (0.0051) |
| Rating C | 0.0628*** | 0.0548*** | 0.0352*** | 0.0672*** |
| | (0.0025) | (0.1157) | (0.0054) | (0.0228) |
| Rating D | 0.0990*** | 0.0964 | 0.0845*** | 0.0765*** |
| | (0.0036) | (0.1205) | (0.0246) | (0.0338) |
| Rating E | 0.1347*** | 0.1326* | 0.1165*** | 0.1288*** |
| | (0.0046) | (0.1079) | (0.0112) | (0.0212) |

| | | | | |
|---|---|---|---|---|
| Rating F | 0.1561***<br>(0.0061) | 0.1958*<br>(0.1053) | 0.1763***<br>(0.0242) | 0.1670***<br>(0.0208) |
| Rating G | 0.1910***<br>(0.014) | 0.2562***<br>(0.0760) | 0.2159***<br>(0.032) | 0.2301***<br>(0.0798) |
| Constant | 0.0919***<br>(0.0104) | 0.2342 ***<br>(0.0907) | 0.0612***<br>(0.0283) | |
| Year F.E. | Yes | Yes | Yes | |
| Observations | 725,641 | 131,385 | 594,256 | |
| $R^2$ | 0.5457 | 0.4321 | 0.5487 | |
| Log likelihood | -324821.11 | -333241.28 | -356752.55 | |

## Table 7 Regression output 3

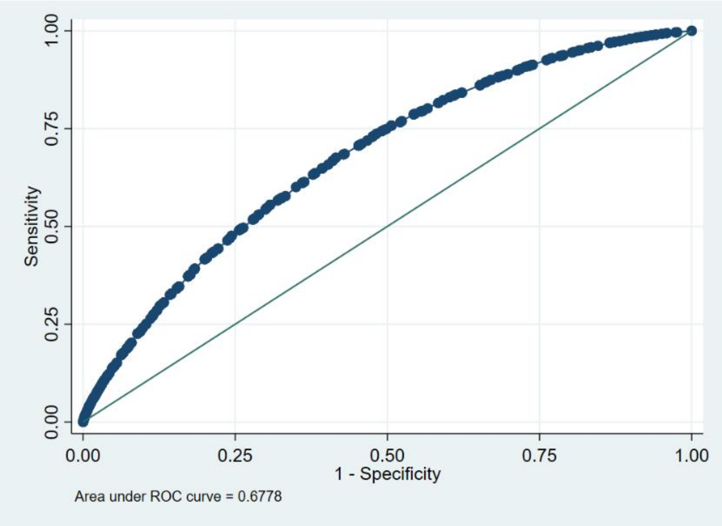
This table exhibits the results of the logistic regression output, which analyzes the relationship between *Default*, the independent variables, and an interaction term is added. Fixed effects are added in the regression as shown. Specification 1 demonstrates the regression coefficients of the baseline model without the interaction term, and specification 2 incorporates the interaction term. In specification 3, the coefficients are indicated where only defaulted loans are analyzed. Specification 4 shows the coefficients where only repaid loans are included. Specification 5 indicates the AMEs for every variable. Standard errors are clustered at the borrower level and displayed in parentheses, and *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| VARIABLES | Baseline | Full sample | Only default loans | Only repaid loans | Average Marginal Effects |
| Low rating | 0.1700***<br>(0.0010) | 0.1480***<br>(0.0027) | 0.2343***<br>(0.0038) | 0.1387***<br>(0.0003) | 0.1689***<br>(0.0011) |
| *DTI* 0 to 10 | 0.0345***<br>(0.0014) | 0.0307***<br>(0.0016) | 0.0214***<br>(0.0061) | 0.0291***<br>(0.0011) | 0.0171***<br>(0.0014) |
| *DTI* 10 to 15 | 0.0525***<br>(0.0015) | 0.0468***<br>(0.0017) | 0.0532***<br>(0.0071) | 0.0322***<br>(0.0020) | 0.0357***<br>(0.0014) |
| *DTI* 15 to 20 | 0.0743***<br>(0.0016) | 0.0669***<br>(0.0019) | 0.0874***<br>(0.0084) | 0.0378***<br>(0.0062) | 0.0537***<br>(0.0016) |
| *DTI* 20 to 25 | 0.0802***<br>(0.0018) | 0.0779***<br>(0.0027) | 0.1123***<br>(0.0094) | 0.0489***<br>(0.0021) | 0.0752***<br>(0.0019) |
| *DTI* 25 to 30 | 0.0902***<br>(0.0018) | 0.0879***<br>(0.0027) | 0.1371***<br>(0.0121) | 0.0581***<br>(0.0026) | 0.0844***<br>(0.0022) |
| *DTI* 30 to 40 | 0.0915***<br>(0.0091) | 0.0831***<br>(0.0131) | 0.1523***<br>(0.0127) | 0.0671***<br>(0.0042) | 0.0883***<br>(0.0031) |
| *DTI* 40 to max | 0.1061***<br>(0.0107) | 0.0816***<br>(0.0064) | 0.2112*<br>(0.0194) | 0.0829***<br>(0.0078) | 0.0853***<br>(0.0073) |
| *DTI* 0 to 10 * Low rating | | 0.0074***<br>(0.0036) | 0.0412***<br>(0.0009) | 0.0021***<br>(0.0006) | 0.0391***<br>(0.0059) |
| *DTI* 10 to 15 * Low rating | | 0.0174***<br>(0.0036) | 0.0671***<br>(0.0011) | 0.0132***<br>(0.0009) | 0.0891***<br>(0.0043) |
| *DTI* 15 to 20 * Low rating | | 0.0205***<br>(0.0035) | 0.0846***<br>(0.0032) | 0.0172***<br>(0.0012) | 0.1372***<br>(0.0075) |
| *DTI* 20 to 25 * Low rating | | 0.0281***<br>(0.0036) | 0.0933***<br>(0.0034) | 0.0199***<br>(0.0018) | 0.1732***<br>(0.0091) |
| *DTI* 25 to 30 * Low rating | | 0.0333***<br>(0.0038) | 0.0965***<br>(0.0046) | 0.0271***<br>(0.0029) | 0.1956***<br>(0.0122) |
| *DTI* 30 to 40 * Low rating | | 0.0425***<br>(0.0041) | 0.1023***<br>(0.0071) | 0.0465*<br>(0.0329) | 0.2781***<br>(0.0222) |
| *DTI* 40 to max * Low rating | | 0.0499***<br>(0.0183) | 0.1230***<br>(0.0101) | 0.0686*<br>(0.0512) | 0.3387***<br>(0.0458) |
| Constant | 0.102***<br>(0.0011) | 0.126***<br>(0.0012) | 0.2192***<br>(0.0087) | 0.0821***<br>(0.0064) | |
| Year F.E. | Yes | Yes | Yes | Yes | |
| Observations | 725,641 | 725,641 | 725,641 | 725,641 | |
| $R^2$ | 0.045 | 0.066 | 0.1720 | 0.1891 | |
| Log likelihood | -273429.04 | -281840.21 | -279213.81 | -281282.22 | |

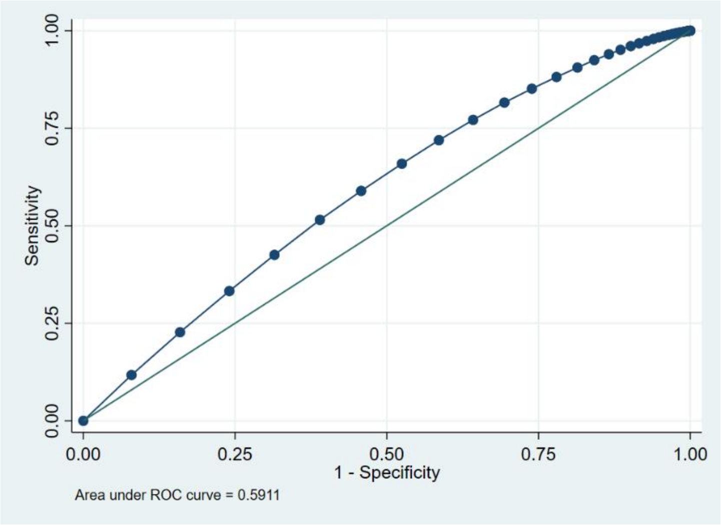
## Graph 1 ROC Curves

Panels (A) and (B) display the ROC curves for the relationship between *Default* and the variables *Interest* and *FICO score*, respectively. Panel (C), (D), and (E) show the relationship between Default and grade, income, and employment, respectively. Panel (F) represents the relationship between Default and all the variables. Panel (G) illustrates the relationship between Default and all the variables in the model of LendingClub.

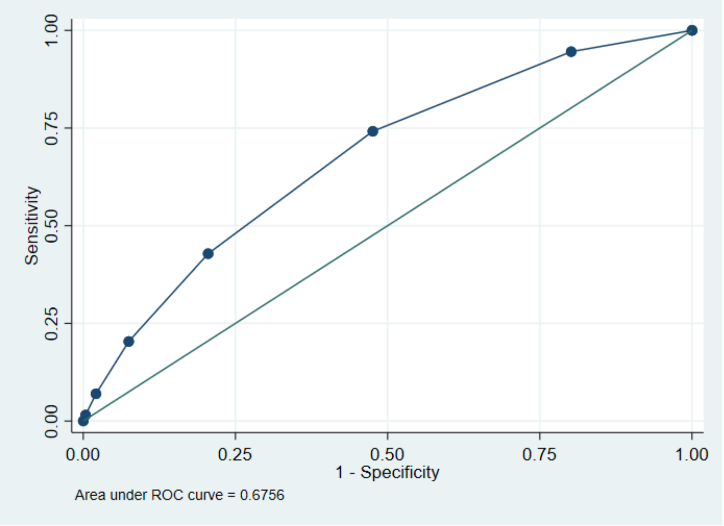
(A) Interest rate



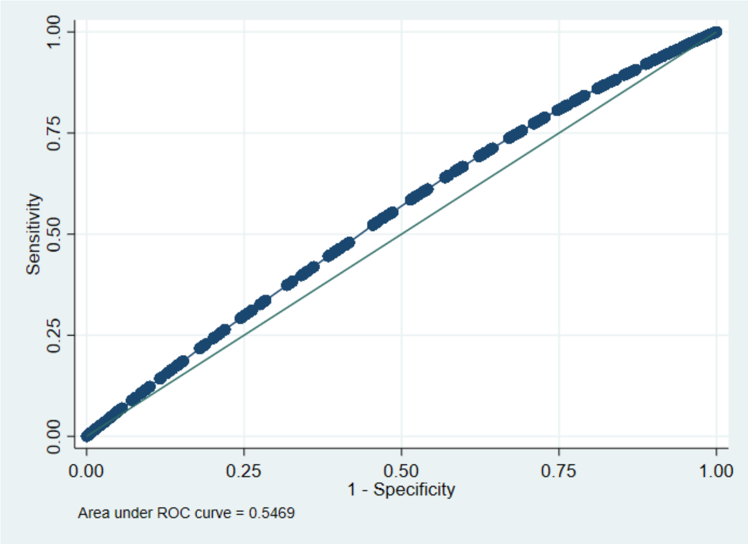
(B) FICO score



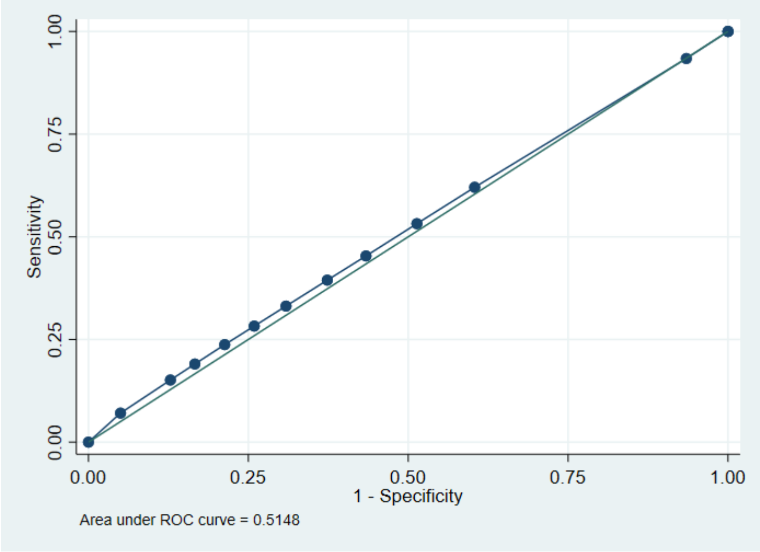
(C) Grade



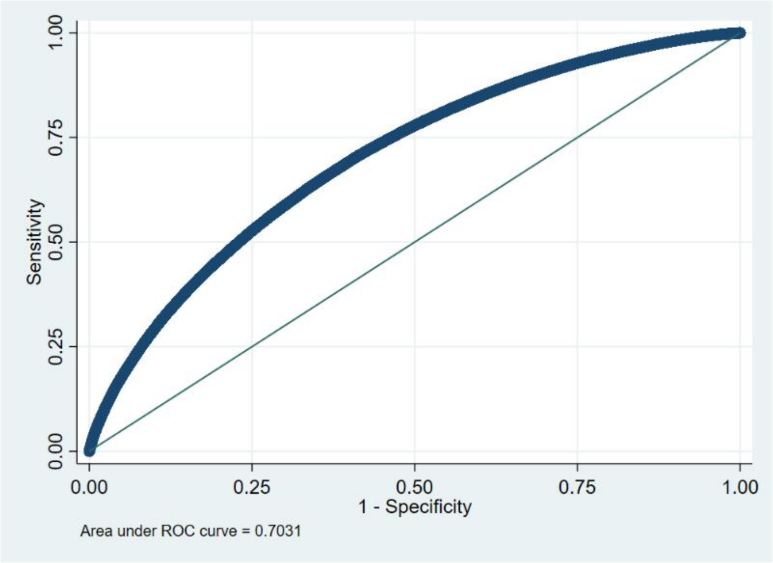
(D) Income variable



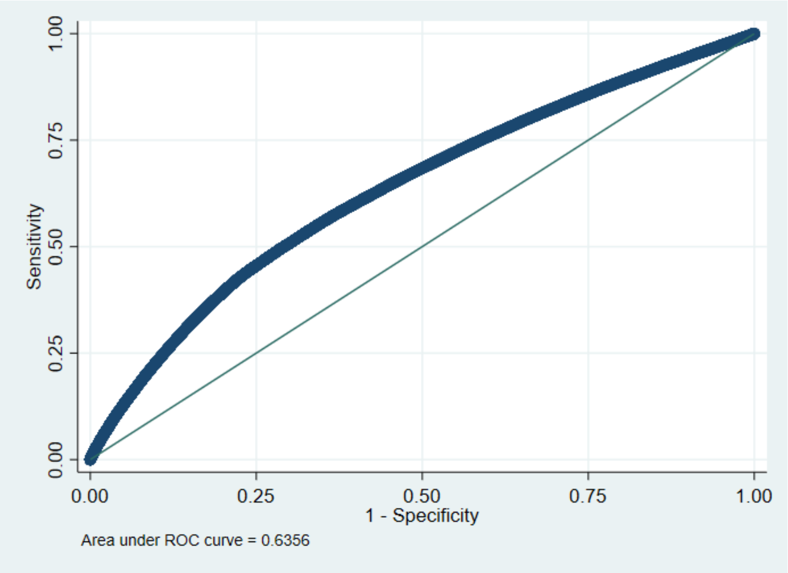
(E) Employment variable



(F) All variables in the conducted model



(G) All variables in LendingClub's model


Sensitivity
1 - Specificity
Area under ROC curve = 0.6356