Peer to Peer Lending Risk Analysis: Predictions from Lending Club

Yueqi Gu[1](✉), Lingqi Guo[2], Chongyue Ma[3], Haoyu Wang[4], and Xiaoran Wei[5]

[1] Statistics, North Carolina State University, Raleigh, USA
ygu5@ncsu.edu
[2] Accounting and Business Analytics, McGill University, Montreal, Canada
lingqi.guo@mail.mcgill.ca
[3] Management, University of Miami, Miami, USA
cxm1446@miami.edu
[4] Pure Mathematics, University of California-Irvine, Irvine, USA
haoyuw16@uci.edu
[5] Applied Statistics and Mathematical Science, University of Toronto, Toronto, Canada
xiaoran.wei@mail.utoronto.ca

Abstract. In this study, we use data findings of a lending club, a p2p company, to visualize, categorize, and use statistical techniques as our research method. In the case of statistical techniques, a combination of logistic regression and random forest is mentioned. The study then analyzes the future risk of the company through two aspects of the lending club: the geographical factors of loan origination and the use of loans. Based on the data, we found that the loans that have the potential to become bad loans are the ones that may lead to a high risk for the lending club in the future. Therefore, with the risk analysis obtained from the data, the lending club needs to anticipate the possibility of bad loans and thus avoid these potential risks.

Keywords: Lending Company · Machine Learning · Risk Analysis · Data Analysis

1 Introduction

This study focused on the loan analysis of Lending Club based in the United States. Lending Club is a peer-to-peer lending company based in the United States. Investors provide funds for potential borrowers, and investors earn a profit depending on their risk (the borrowers' credit score). Lending Club provides the "bridge" between investors and borrowers [8]. Our group divided the whole study into five sections. They are "Data," "Methodology," "Analysis," "Risk Analysis," as well as "Conclusion." The "Data" section talks about the rough range of borrowing. Also, it compared the trend of actual borrow amount between borrower and lender. Moreover, it also illustrated the shape of the overall data on borrowing amount change and the types of data in two notebooks. The "Methodology"

Y. Gu, L. Guo, C. Ma, H. Wang and X. Wei—Contributed equally.

© The Author(s) 2023
Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 750–759, 2023.
https://doi.org/10.2991/978-94-6463-005-3_76

section focused on visualization, categorization, and statistical techniques to simplify complex datasets and draw conclusions. Then, in the "Analysis" section, it displayed different types of loans and the good and bad of loan purposes. More importantly, it also showed the good and bad rates of loans that belong to various job types. The fourth section is "Risk analysis," this paragraph talks about risk analysis based on loan information and other loan properties. Unlike a commercial bank, the platform does not take risks through its contractual positions. Banks accumulate risks by taking classes on their balance sheet, but platforms decentralize the risks by spreading them to their users [4]. The risk analysis is mainly based on state credit, credit rating, and loans that could become non-performing loans. The risk of the Lending Club company may result from the application for the loan applications of bad loans. Credit risk or default risk involves the inability or unwillingness of a customer or counterparty to meet commitments with lending, trading, hedging, settlement, and other financial transactions [2]. Credit Risk is generally made up of transaction, default, and portfolio risk. The portfolio risk, in turn, comprises intrinsic and concentration risk [5]. The credit risk of a bank's portfolio depends on both external and internal factors [8].

The last section is "Conclusion." This paragraph concludes the important elements of previous sections. It also gave the Lending Club company some advice to avoid the risks they will confront in the future. We also argue that the full development of the sector requires much further work addressing the risks and business and regulatory issues in P2P lending, including risk communication, orderly resolution of platform failure, control of liquidity risks, and ministration of fraud, security, and operational risks [4]. This will depend on developing a reliable business process, the promotion of as complete a degree as possible of transparency and standardization, and appropriate regulation that serves customers' needs [2]. The main topic of this paper is the analysis and the description of the study. We find that credit grade, debt-to-income ratio, FICO score, and revolving line utilization play an important role in loan defaults. Loans with lower credit grades and longer duration are associated with a high mortality rate [5]. Accordingly, this will make sense for the result, which will be talked about next. The result is consistent with the Cox Proportional Hazard test, which suggests that the hazard rate or the likelihood of the loan default increases with the credit risk of the borrowers [6]. Under this circumstance, The Lending Club must find ways to attract high FICO scores and high-income borrowers to sustain their businesses [1].

2 Data Description

As the most basic and critical condition, data is the minimum prerequisite for every excellent analysis. Lending Club is an online peer-to-peer loan platform; the data they care about most should relate to all fields about borrowers. The dataset used in this article contains 890 thousand observations of the lending company from 2007 to 2015. However, not all features are valuable and crucial in this analysis. The author eliminates several aspects from the dataset, such as the borrower's zip code, member ID, and those fields that contain a larger number of missing values. Loan status, purposes, amounts, and credit scores are the main features the author emphasizes in the paper. The data type of this set is both qualitative and quantitative, which contains all discrete, continuous, ordinal, and nominal values.

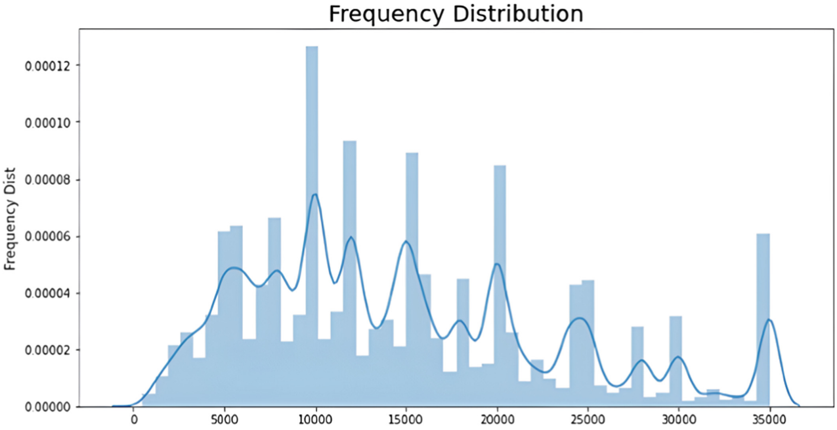

Fig. 1. The frequency distribution of borrower's loan amount

According to Fig. 1, the multimodal histogram, 10 thousand to 20 thousand, is usually the most common range people are willing to borrow online. The maximum loan quantity the company is capable offer is 35 thousand. Moreover, the dataset also introduces various loan purposes. When these purposes are combined with the loan amount borrowed and repaid, the lending company will be given a glimpse into the borrower's life and be more aware of the potential risks. Debt consolidation, credit card refinancing, and home improvement are the three most used reasons for borrowing money. In addition, educational, small business, and renewable energy relate purposes are the top three fields with the worst repayment rate.

Figure 2 shows three multimodal histograms that compare the movement between loans among borrowers, investors, and the actual funded amount. Obviously, the trend of these three histograms is very similar, which gives us an idea that the investors are willing to lend borrower loans as the amount borrowers want when they meet the criteria. The criterion for borrowing is stringent in the Lending Club; the borrower is likely not to get the full applied loan if they fail to pass the final credit report [6].

From Fig. 3, besides showing 2015 was the year with the highest average loan amount, a prominent rising trend is also displayed in the chart. It is a sign that indicates the Lending club and investors have increased their average loan amount each year. This phenomenon represents a slow economic recovery in the United States at that time. Recovery in economics is a good sign for all lending companies to keep them away from the brink of bankruptcy [1].

3 Methodology

The research mainly tackles the problem using visualization, categorization, and statistical techniques to enhance the interpretability of complex datasets and draw conclusions based on features that are correlated.

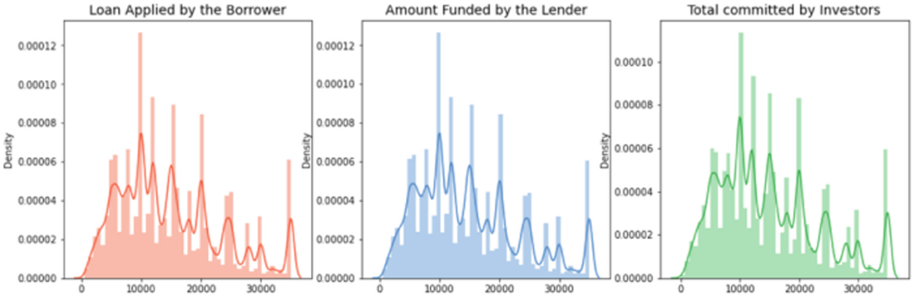

Fig. 2. The frequency distribution of loans between borrower applied, lender, funded, and final issued

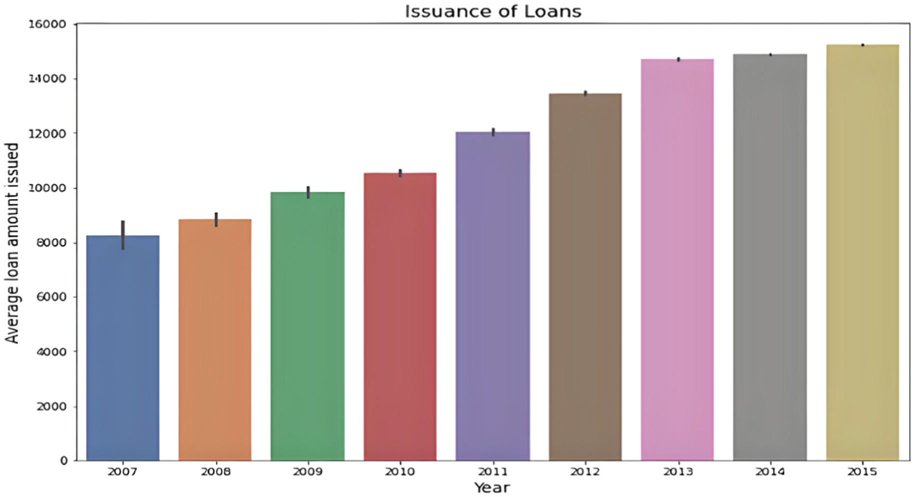

Fig. 3. The average loan amount issued from 2007 to 2015 in Lending Club

3.1 Visualization

Data visualization has become an integral aspect of today's decision-making process, in which raw data is considered and then turned into graphs, charts, and infographics to offer a visual depiction of data. It translates complex data into tangible presentations of information.

The research first used frequency distribution density graphs, box plots, and violin charts, which all focus on the distribution of numerical data and compares the distribution amongst various groups. The graphs present not only a macro-observation but also crucial elements such as mean, median, outlier, maxima, and minima. Other than distribution diagrams, heatmaps were utilized as well. Heatmaps are quite beneficial in terms of giving a broad perspective of numerical data rather than extracting single data points. Besides, pie charts and line plots are used as well to observe the portion distribution and tendency in a complicated dataset.

Other than graphs that focus mainly on numeric values, choropleth maps are presented to depict geographically split zones or regions by coloring in response to their value. It enables studies about how a variable evolves over an area. However, it is quite known that data visualization could have some drawbacks at the same time. For instance, visualization tends to be based more on estimations rather than accurate analysis. The change of scale and weight could easily lead to speculative results. What's more, data visualization can be biased and fail to convey sufficient information due to misconduct of designing and interpreting. The use of data visualization ought to be legible, precise, and concise.

3.2 Categorization

In addition to visualization, the analysis also implements categorization when analyzing the data. For example, debts were divided into good debts and bad debts, whereas income was classified into low income, medium income, and high income. Data categorization separates data into subsets that share similarities in certain aspects. It is different from classification, which assigns data into different unique classes within a system. The advantage of grouping is that it can save the volume of computation needed and draw conclusions on a feature basis. Nevertheless, categorization requires analysts' personal judgment based on experience and sensitivity to the datasets. Thus, it could result in inaccuracy and prejudices, which can potentially affect the results significantly.

3.3 Statistical Techniques

The research used the above methods in the study, combining them with statistical approaches such as logistic regression and random forest. Logistic regression is a statistical model that is primarily used for modeling binary dependent variables, with many more complex extensions existing (Juliana, Meurer 2016). It is simple to implement and effective while having shortcomings such as poor performance with non-linear data and highly correlated features. Random forest is a machine learning method for classification, regression, and other tasks by constructing a multitude of decision trees at training time [3]. It can operate with a massive volume of data while reducing errors and the impact of outliers. However, random forest is hard to interpret with meanings, and it won't work unless features have some predictive powers.

4 Loans Type, Quality, Analysis

The research analyzes the loan quality problems faced by the company, including numbers and quality of loans, and analyzes the existing risks.

4.1 Loan Quality and Purpose Distribution

For the first step, the author presented several types of loans for years (Fig. 4). The largest number of loan statuses is "Current" in the total count. In subsequent years, the number of loan status is gradually increasing.

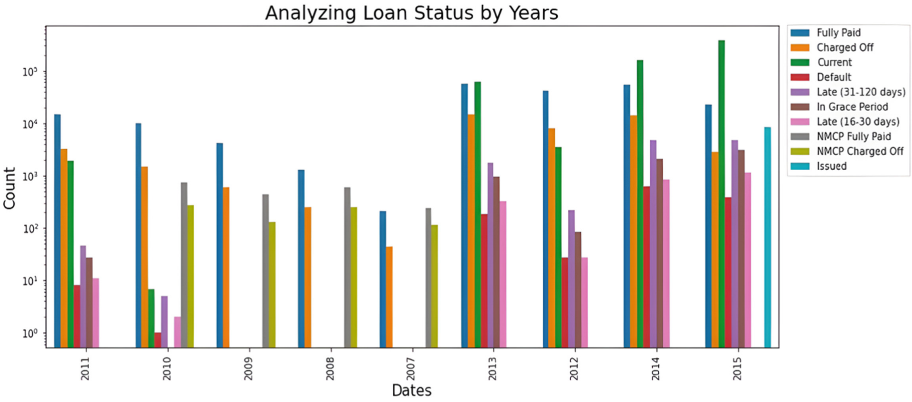

Fig. 4. 10 types of loans from 2011 to 2015

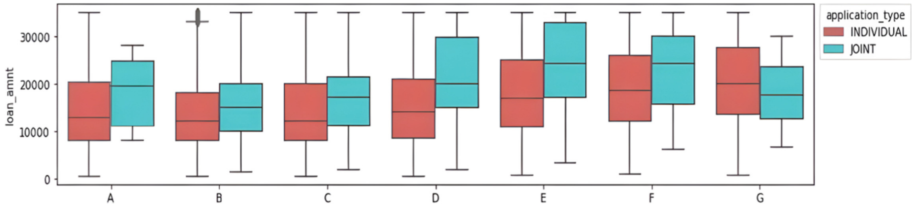

Fig. 5. Application type and grade of loans

The number of loan status is in the majority for individuals compared to joint (Fig. 5). The main purpose in individuals is debt consolidation. However, in grades A–F, almost every type of loan is larger for the joint than for the individual.

Good debt has the possibility of increasing individual or collective net worth or improving lives in essential ways, such as student loans and business loans. Bad debt includes borrowing money to buy assets that depreciate quickly or for consumption only and does nothing to improve individual or collective financial situations. Current loans still have the possibility of becoming bad loans in the future. The figure below shows how we get the average annual income (Fig. 6).

4.2 Factors of Bad Loans

Loans issued by region to observe regional tendencies that will help us figure out which region provides Lending Club. The regions with the most loans issued were the South-East, West, and North-East. Beginning in 2012, debt-to-income ratios in the West and South-West increased rapidly. Interest rates fell rapidly in the West and South-West. This could explain the rise in the debt-to-income ratio.

The Midwest and South-East regions had the most defaults in the previous year. Loans with a high-interest rate (more than 13.23%) are more likely to default. Loans with a longer maturity period (60 months) are more prone to default. There has to be

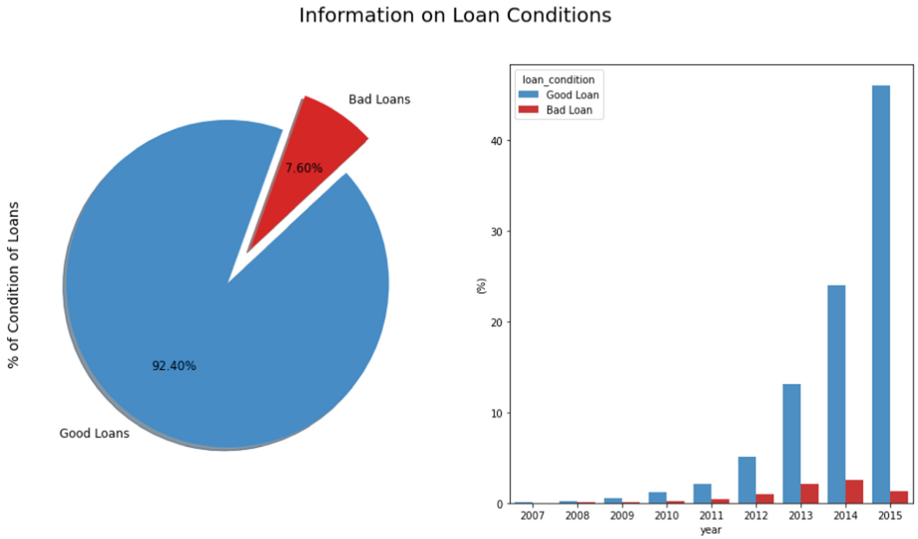

Fig. 6. Percentage of Good Loans and Bad Loans

a better risk assessment using Average Interest Rates Loan Status Distribution. The majority of loan statuses are Current, but some loans may become defaulted.

4.3 Risk Analysis

The data used in this study is derived from the loan origination information and other loan attributes given in the lending club's python notebook. The loan origination information includes the credit scores, employment, income, loan purpose, geographic location, and interest rate of the borrower [7]. The other loan attributes include the percentage of good loans and bad loans in different regions and for different loan purposes. The company's primary lending range is between $0 and $35,000.

This study analyzes the risk profile of a lending club, a p2p company, based primarily on these data. First, from the loan status, the largest number of loan status is "Current" in the total count. In subsequent years, the number of loan status is gradually increasing, like loan status "Default", which peaked in number in 2014 from 2010 to 2015.

In addition, borrowers are categorized into seven credit scores, A through G. The lower the credit score, the higher the risk to the investor [7]. In Fig. 7, we find that those with lower scores receive more loans (which may lead to higher levels of risk). Logically, the lower the score, the higher the interest rate paid by the client to the investor. Interestingly, however, customers with a "C" score were more likely to default on their loans. So, predicting in advance which loans are likely to become bad is good for the lending club's future growth.

Of all the bad loans, the one we are most interested in is the defaulted loans. From Fig. 8, loans with higher interest rates (above 13.23%) are more likely to become bad loans. Loans with longer terms (60 months) are more likely to become bad loans. Too many loans becoming bad will lead to higher lending risk for the lending club, thus

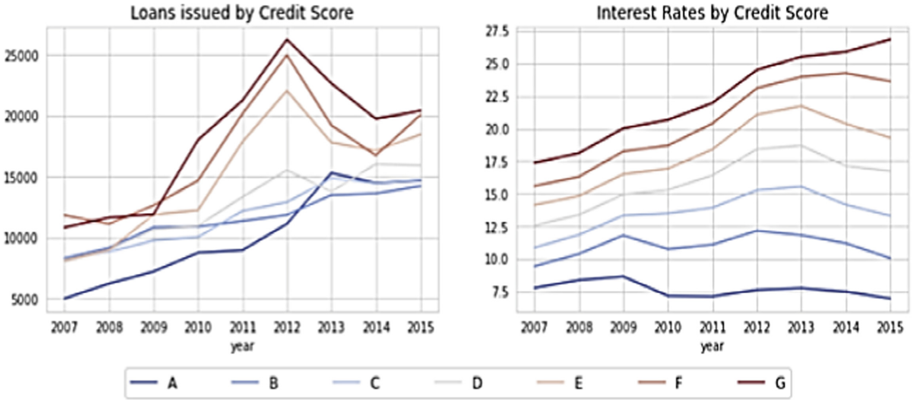

Fig. 7. Loans issued and Interest Rates by Credit Score

not attracting more investors to invest. In this way, the capital chain will be difficult to maintain and may even fall into bankruptcy.

The data information on the purpose of the loan shows that those who apply for education and small business tend to have a higher risk of becoming bad loans. And the most common reason customers apply for loans is debt consolidation. From the distribution of purposes, we can likewise find that individuals make up the majority of loan statuses, with the highest number of loans used for debt consolidation compared to joint loans. Although the number of individual loans is greater than the number of joint loans, the amount of joint loans is greater than the number of individual loans in almost all loan types in scores A–F, and only in score G is the number of individual loans greater. Also, in scores D, E, and F, the amount of joint loans is higher than the other scores, with a median of more than 20,000. Therefore, we can know that the amount of the joint loan is usually higher than the amount of the personal loan. From the currently available data, we cannot determine which is riskier for the lending club, the joint loan, or the individual loan. But what the lending club needs to pay attention to when avoiding the risk is to reduce more loans to become bad loans, which will avoid many risks.

5 Conclusion

To conclude, our group did some research on Lending Club Loan Analysis. The article used visualization and categorization as a method to analyze the risk of the company. Lending Club is a peer-to-peer lending company, where its head office is located in San Francisco, California. It was the first peer-to-peer lender to register with the Securities and Exchange Commission to issue securities and trade loans in the secondary market. As we can see from Wikipedia, Lending Club is a loan company. The company's core product is unsecured personal loans. The borrowers use to raise car loans and cards. However, Lending Club needs partner banks to provide the credits. This may cause pressure on its profitability. In the study, the data type of the set which appeared at the beginning of the paragraph is both qualitative and quantitative. The method this study used is visualization, categorifications as well as statistical techniques. After that, the

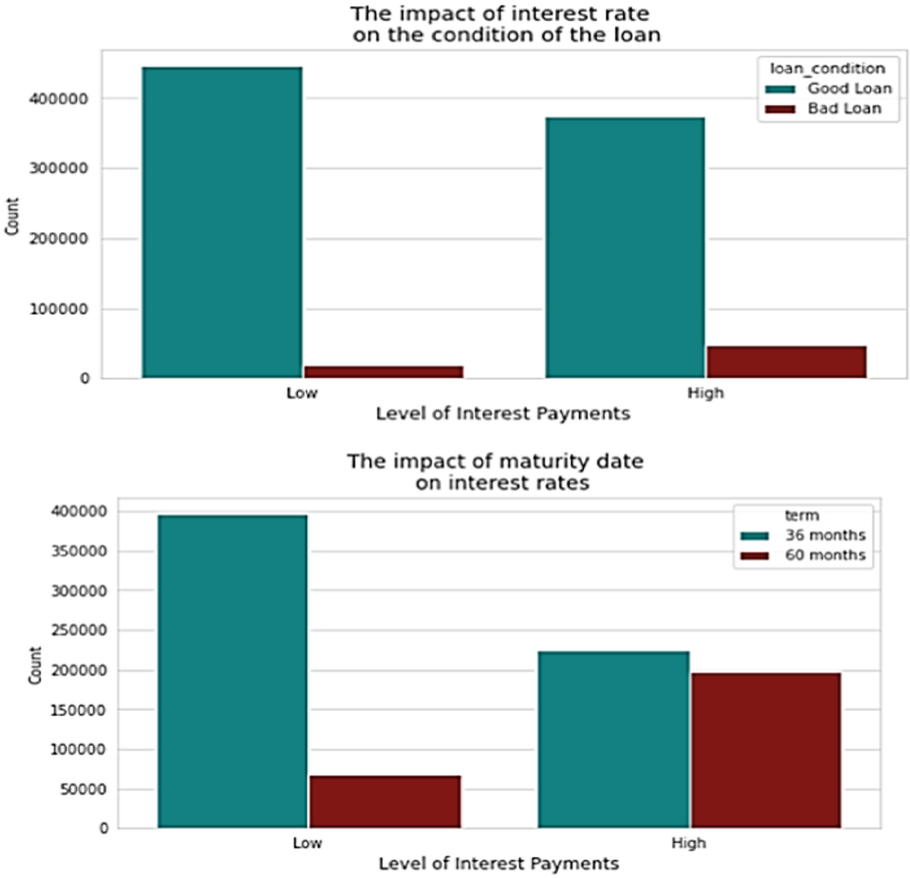

Fig. 8. Different Loan Conditions with High and Low-Interest Rates

study described several loan types, quantities as well as analyses. Readers could see from the chart as well as the models this study applied. Eventually, the last part of the body paragraph emphasized the risk analysis.

In a nutshell, this study mainly talks about various aspects as well as risk analysis of the loan. The study analyzed bad loans from various aspects. A bad loan is the main element at the beginning of the paragraph. The study used different charts as well as icons to help readers understand the topic better. The most important thing the company needs to pay attention to is that some current loans still can become bad loans in the future. They need to try their best to reduce this potential risk. If they don't pay attention to current loans, they will have a great risk in the future of the company.

The suggestion that our group would like to give to the Lending Club Loan company is that focus on current loans as well. In the future, they could predict the possibility of current loans which had the possibility to become bad loans. If they could predict this element, they will avoid many problems in the future. Under this circumstance, the company will have a better effect on loans. More importantly, they will make more profits in the future.

References

1. Didier T, Huneeus F, Larrain M, Schmukler SL (2021) Financing firms in hibernation during the COVID-19 pandemic. J Financial Stab 53:100837
2. Emekter R, Tu Y, Jirasakuldech B, Lu M (2014) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending, pp 54–70. tandfonline.com/doi/abs/10.1080/00036846.2014.96222210.1080/00036846.2014.962222
3. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol 1, pp 278–282
4. Lenz R (2017) Peer-to-peer lending: opportunities and risks. cambridge.org/core/journals/european-journal-of-risk-regulation/article/abs/peertopeer-lending-opportunities-and-risks/9B9E21667A148330DDA491775A23AF5E
5. Milne A, Parboteeah P (2016) The business models and economics of peer-to-peer lending. papers.ssrn.com/sol3/papers.cfm?abstract_id=2763682
6. Nowak A, Ross A, Yencha C (2018) Small business borrowing and peer-to-peer lending: evidence from lending club. Contemp Econ Policy 36(2):318–336
7. Reddy S, Gopalaraman K (2016) Peer to peer lending, default prediction-evidence from lending club. J Internet Bank Commer JIBC 21(3):1
8. Spuchľáková E, Valašková K, Adamko P (2015) The credit risk and its measurement, hedging and monitoring. Procedia Econ Financ 675–681. https://doi.org/10.1016/S2212-5671(15)00671-1
9. Tolles J, Meurer WJ (2016) Logistic regression: relating patient characteristics to outcomes. JAMA J Am Med Assoc 316(5):533–534

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.