IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006232|

Prediction of Loan Strategy for Lenders with Apache Spark

N.S. ADITHYAN, RAHUL.C, YERRASURA SUMANTH, R.RAJA

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

Assistant Professor, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT: In the past few years, Peer-to-Peer lending (P2P lending) has grown rapidly in the world. The main idea of P2P lending is disintermediation and removing the intermediaries like banks. For a small business and some individuals without enough credit or credit history, P2P lending is a good way to apply for a loan. However, the fundamental problem of P2P lending is information asymmetry in this model, whichmay not correctly estimate the default risk of lending. Lenders only determine whether or not to fund the loan by the information provided by borrowers, causing P2P lending data to be imbalanced datasets which contain unequal fully paid and default loans. Imbalanced datasets are quite common in the real worlds, such as credit card fraud in transactions, bad products in the plant and so on. Unfortunately, the imbalanced data are unfriendly to the normal machine learning schemes. In our scenario, models without any adaptive methods would focus on learning the normal repayment. However, the characteristic of the minority class is critical in the loaning business. In this study, we utilize not only several machine learning schemes forpredicting the default risk of P2P lending but also re-sampling and cost-sensitive mechanisms to process imbalanced datasets. Furthermore, we use the datasets from Lending Club to validate our proposed scheme. The experiment results show that our proposed scheme can effectively raise the prediction accuracy for default risk.

I. INTRODUCTION

Peer-to-Peer (P2P) lending, consists of the project of matching anonymous lenders with borrowers through an electronic platformso lenders could directly invest on (lend to) certain borrowers .In general, lenders could earn higher returns relative to savingsand other investment products offered by banking when borrowerspay back their loans as scheduled. However, the loans on the P2Pmarket are unsecured and investors need to tolerate the risk oflosing part or even all of their principal if borrowers default theloans. To help investors find out the safer loans with the relativelylower risk, it is beneficial to evaluate each loan from the perspectiveof “the risk level", which is typically done by estimating the probability of default (PD). Loans with lower PDs are consideredsafer than those with higher PDs and vice versa. The PD for eachloan can be predicted by considering its characteristics, such asthe loan amount, the loan purpose, the assets of the borrowers,etc. The above-mentioned approach is known as the credit scoringapproach, which poses a classification problem that classifies theloans into either (1) the default case if the predicted PD exceeds a certain predefined threshold, or (2) the non-default case otherwise.Subsequently, the credit scoring approach recommends lenders toinvest in non-default loans or the loans with lower predicted PDsbecause of the potentially lower risk.

II. EXISTING SYSTEM

The credit scoring systems mainly focus on loan default probability. By analysing borrower’s interest rate and lenders’ profitability, the results indicate that the P2P lending is not a trend in current market. By doing so, this method can simplify optimization problem to an integer linear programming. Differ from other studies, this research estimates expected profitability in other metrics, such as annualized rate of return (ARR). The metrics used in estimation are designed on the basis of an imbalanced dataset. Although there were some researches in prediction P2P lending default risk, they did not focus on addressing the problem that imbalanced datasets bring. Regularly, there are two classes in imbalanced datasets like the majority of negatives class and the minority of positive class. These types of data presume an issue for data mining since standard classification algorithms normally consider a balanced training set and this supposes a bias towards the majority (negative) class metrics that they used was accuracy which was not suitable for imbalanced datasets. Existing

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006232|

classification algorithms are poor performance in imbalanced datasets.The sampling strategies and cost sensitive learning to address the issue of expectation imbalanced datasets.

III. PROPOSED SYSTEM

The P2P lending datasets contain many attributes whichare empty for most records. Therefore, we remove theseattributes and modify the nominal features by using one-hot-encoding technique that can transform nominal features to bea format suitable for classification. For instance, we have afeature ‘‘purpose of the loan’’ which has string value suchas ‘‘Car’’, ‘‘Business’’, and ‘‘Wedding’’. Normally, we useordinal value to encode these to be numbers such as 0,1, and 2. However, in machine learning schemes, differentcategories have the same weight. Thus, the ordinal technique cannot be implemented in machine learning becausethe lowest and the highest value will affect the classificationresult. One-hot-encoding uses one Boolean column for eachcategory which has different weight. We use libsvm library to convert the features into encoded format. Then the dataset is sent to training and prediction with analytics.

IV. LITERATURE SURVEY

1]Project Title: CLASSIFICATION OF IMBALANCED DATA: A REVIEW
Author Name : YANMIN SUN, ANDREWK. C. WONG, MOHAMED S. KAMEL
Year of Publish: 2009
Abstract:
Classification of data with imbalanced class distribution has encountered a significant drawback of the performance attainable by most standard classifier learning algorithms which assume a relatively balanced class distribution and equal misclassification costs. This paper provides a review of the classification of imbalanced data regarding: the application domains; the nature of the problem; the learning difficulties with standard classifier learning algorithms; the learning objectives and evaluation measures; the reported research solutions; and the class imbalance problem in the presence of multiple classes.

2]Project Title : Determinants of Default in P2P Lending
Author Name : Carlos Serrano-Cinca☯, Begoña Gutiérrez-Nieto*☯, Luz López-Palacios☯
Year of Publish: 2015
Abstract:
This paper studies P2P lending and the factors explaining loan default. This is an important issue because in P2P lending individual investors bear the credit risk, instead of financial institutions, which are experts in dealing with this risk. P2P lenders suffer a severe problem of information asymmetry, because they are at a disadvantage facing the borrower. For this reason, P2P lending sites provide potential lenders with information about borrowers and their loan purpose. They also assign a grade to each loan. The empirical study is based on loans' data collected from Lending Club (N = 24,449) from 2008 to 2014 that are first analyzed by using univariate means tests and survival analysis. Factors explaining default are loan purpose, annual income, current housing situation, credit history and indebtedness. Secondly, a logistic regression model is developed to predict defaults. The grade assigned by the P2P lending site is the most predictive factor of default, but the accuracy of the model is improved by adding other information, especially the borrower's debt level.

3]Project Title: Credit risk evaluation in peer-to-peer lending with linguistic data transformation and supervised learning
Author Name : J´ozsefMezei, Ajay Byanjankar3, Markku Heikkil¨
Year of Publish: 2018
Abstract:
The widespread availability of various peer-to-peer lending solutions is rapidly changing the landscape of financial services. Beside the natural advantages over traditional services, a relevant problem in the domain is to correctly assess the risk associated with borrowers. In contrast to traditional financial services industries, in peerto- peer lending the unsecured nature of loans as well as the relative novelty of the platforms make the assessment of risk a difficult problem. In this article we propose to use traditional machine learning methods enhanced with fuzzy set theory based transformation of data to improve the quality of identifying loans with high likelihood of default. We assess the proposed approach on a real-life dataset from one of the largest peer-to-peer platforms in Europe. The results

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006232|

demonstrate that (i) traditional classification algorithms show good performance in classifying borrowers, and (ii) their performance can be improved using linguistic data transformation.

4]Project Title: Handling Imbalanced Data: A Survey
Author Name : Neelam Rout, Debahuti Mishra and Manas Kumar Mallick
Year of Publish: 2018
Abstract:
Nowadays, handling of the imbalance data is a major challenge. Imbalanced data set means the instances of one class are much more than the instances of another class where the majority and minority class or classes are taken as negative and positive, respectively. In this paper, the meaning of the imbalanced data, examples of the imbalanced data, different challenges of handling the imbalanced data, imbalance class problems and performance analysis metrics for the imbalanced data are discussed. Then different methods are summarized withtheir pros and cons. Finally, the examples of the imbalanced data sets having low-to-high imbalance ratio (IR) values are shown.

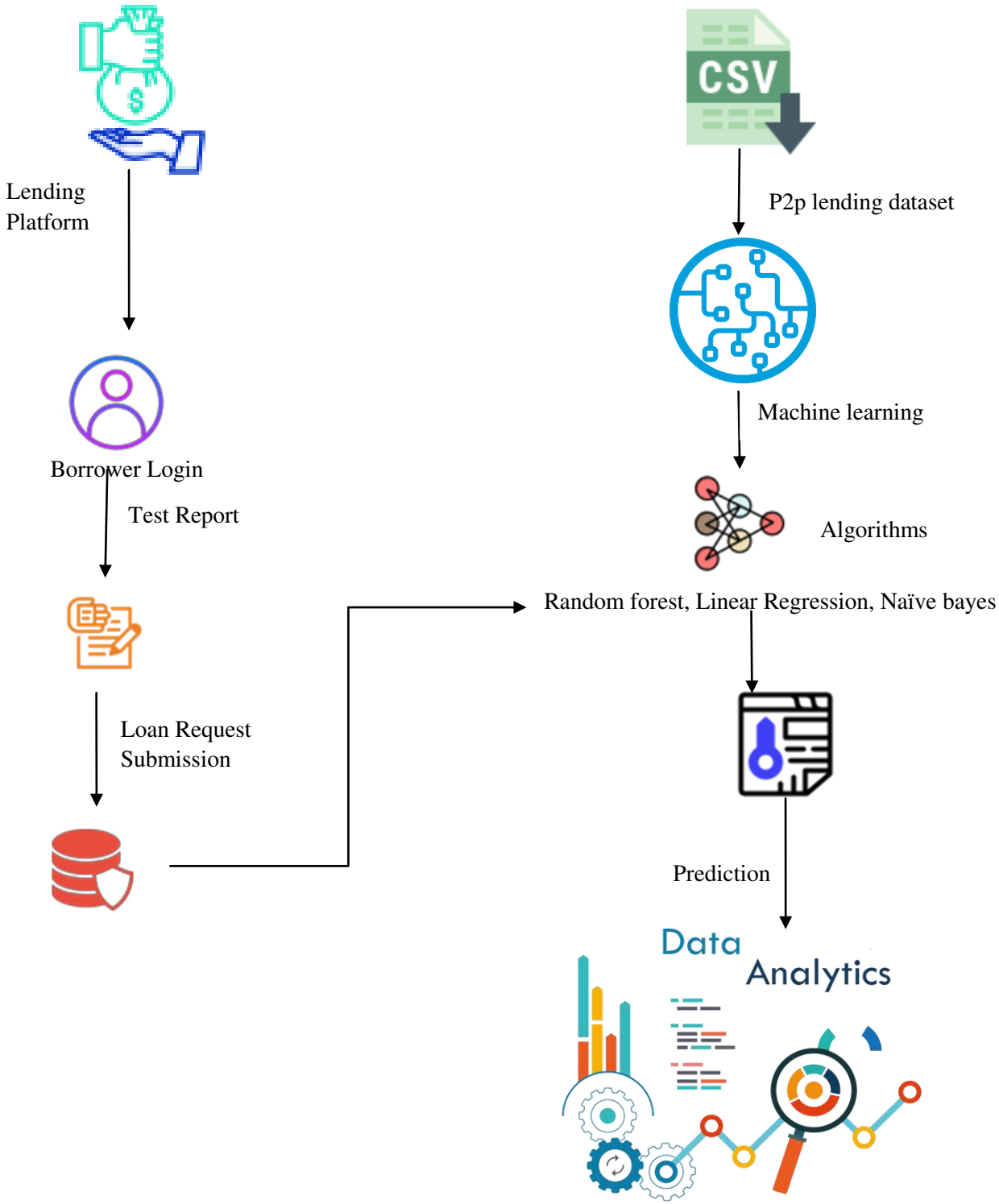
5]Project Title: Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data
Author Name : Jei Young Leea
Year of Publish: 2020
Abstract:
Usingdata from Lending Club, we analyzed funded loans between 2012 and 2013, the default status of which were mostly known in 2018. Our results showed that both the borrower characteristics and the conditions of the loan were significantly associated with the loan default rate. Results also showed that the sentiment of a user-written loan description influenced the borrower's loan interest rates. It contributes to expanding the scope of peer-to-peer (P2P) loan research by implementing unstructured data as a new model variable. Financial counselors need to consider the growth potential of the P2P loan market using data analysis.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006232|

Architecture Diagram:



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006232|

V. CONCLUSION

P2P lending companies may bear less transaction costs than conventional financial institutions do, since its business model is simpler: they do not capture deposits, they are not under strict banking regulations, they do not maintain idle balances; they just put borrowers in contact with lenders. Besides, this is done by means of an online platform where most of the processes are automatized. Operating cost is the most important factor explaining interest margins in banking, and P2P lending platforms–like other online businesses- have the use of technologies as strength. This can lead to improving the efficiency, a very important factor in a market where money is bought and sold. Money is a non-differentiated product and its price, the interest rate, is what matters most. P2P lending can alleviate credit rationing, especially for thoseborrowers placed in the long tail of credit. These advantages could explain P2P lending growth,but it is not problem-free. In the banking business model, the credit risk is assumed by thefinancial institution, which has risk management departments with skilled financial analysts,supposedly more expert than individual lenders.The paper analyzes whether the information provided by the P2P lending site, a grade thatqualifies the loan, complemented with loan and borrower characteristics, explains loan defaultsand reduces information asymmetry. Firstly, a hypotheses test and a survival analysis havebeen performed on the factors explaining loan defaults. In this paper,different types of existing techniques are discussed for tackling the imbalance classproblems but still improvement techniques are needed, necessarily. It is also knownthat the ensemble learning algorithms are the useful and powerful methods to dealwith the imbalance class problem. Some of the imbalanced data sets have beenshown with different IR values in the tabular manner. It is very important to balance the imbalance data with effective techniques and at the same time, cost factorshould be given attention.

REFERENCES

1. He, Habib, and Edwardo Garcia. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21 (9): 1263–1284.
2. Van Pulse, Jason, and Tag hi Jehoshaphat. 2009. Knowledge Discovery from Imbalanced andNoisy Data. Data and Knowledge Engineering 68 (12): 1513–1542.
3. Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 42 (4): 463–484.
4. He, Habib, and Yunnan Ma. (eds.). 2013. Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley.
5. Yang, Anglia, and Wu Donning. 2006. 10 Challenging Problems in Data Mining Research. International Journal of Information Technology and Decision Making 5 (04): 597–604.
6. Wang, Shu, and In Tao. 2012. Multi Class Imbalance Problems: Analysis and Potential Solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42 (4): 1119–1130.
7. Lakshmi, T. Jay, and C. Pradesh. 2014. A Study on Classifying Imbalanced Datasets. In First International Conference on Networks and Soft Computing (ICNSC), IEEE.
8. European Commission. Crowdfunding in the EU Capital Markets Union,Report, Commission Staff working document. 2016.
9. Mo S, Chen KC, Ye C. The Evolving Role of Peer-to-Peer Lending: A New Financing Alternative. J Int Acad Case Stud. 2016 May 1;22(3).
10. Ziegler T, Shneor R, Garvey K, Wenzlaff K, Yerolemou N, Rui H, et al. Expanding Horizons: The 3rd European Alternative Finance Industry Report.NY: University of Cambridge; 2018. p. 23.
11. Milne A, Parboteeah P. The business models and economics of peer-to-peer lending. Euro Credit Res Inst. 2016.
12. Financial Stability Board. Committee on the Global Financial System. 2017 May 22.
13.Hagiu A. Strategic decisions for multisided platforms. MIT. 2014 Jan 1;55(2):71.
14. Van Alstyne M, Parker G. Platform Business: From Resources to Relationships. GfK Market Intel Rev. 2017 May 1;9(1):24-29.
15.Parmentier G, Gandia R. Redesigning the business model: from one-sided to multi-sided. Journal of Business Strategy. 2017 Apr 18;38(2):52-61.

BIOGRAPHY

N.S. ADITHYAN is a B.E. final year student in department of Computer Science and Engineering Department in Velammal Institute of Technology, Panchetti. His current research focuses on Prediction of Loan Strategy for Lenders with Apache Spark.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006232|

RAHUL.C is a B.E. final year student in department of Computer Science and Engineering Department in Velammal Institute of Technology, Panchetti. His current research focuses on Prediction of Loan Strategy for Lenders with Apache Spark.

YERRASURA SUMANTH is a B.E. final year student in department of Computer Science and Engineering DepartmentinVelammal Institute of Technology, Panchetti. His current research focuses on Prediction of Loan Strategy for Lenders with Apache Spark.

R.RAJA is a M.E., Assistant Professor of Computer Science and Engineering Department in Velammal Institute of Technology, Panchetti.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details