ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse99852019.pdf
https://doi.org/10.30534/ijatcse/2019/99852019

Lending Club Default Prediction using Naïve Bayes and Decision Tree

Mogi Jordan Christ [1], Rahmanto Nikolaus Permana Tri[2], Wiranto Chandra[3], Tuga Mauritsius[4]

[1]Information Systems Management Department, BINUS Graduate Program-Master of Information Systems Management,Bina Nusantara University,Jakarta, Indonesia, 11480,
jordan.mogi@binus.ac.id
[2]Information Systems Management Department, BINUS Graduate Program-Master of Information Systems Management,Bina Nusantara University,Jakarta, Indonesia, 11480,
nikolaus.rahmanto@binus.ac.id
[3]Information Systems Management Department,BINUS Graduate Program-Master of Information Systems Management,Bina Nusantara University,Jakarta, Indonesia, 11480,
chandra.wiranto@binus.ac.id
[4]Information Systems Management Department,BINUS Graduate Program-Master of Information Systems Management,Bina Nusantara University,Jakarta, Indonesia, 11480,
tmauritsus@binus.edu

ABSTRACT

Currently P2P lending is one of the most emerging disruptors in the financial sector. Lending Club is a P2P platform based in America. Besides its flexibility to give instant lending this industry have high risk for their investors to lending money. In order to mitigate this risk, this study aims to predict the default risk using decision tree J48 and naive bayes. One of the results in this research show that J48 and Naïve Bayes are both good in predicting the default in P2P lending sector. Another contribution of the paper might be useful for similar companies to see which factors that influence the most to default loan status.

Key words: Lending Club, Naive Bayes, Default, Decision Tree, J48

1. INTRODUCTION

Peer-to-peer (P2P) lending are money lending activities directly from each individuals without intermediaries from financial institutions [1]. Because of this, individuals allowed to lend and borrow directly on an Internet-based platform, without the involvement of financial intermediaries. In P2P Lending, borrowers apply for lending, called listings, by determining loan details, such as the loan amount and description. Then, potential lenders are permitted to take part in determining the amount of the loan they will give. If the total dollar amount requested by a list is fulfilled within a specified time period, the list becomes a loan [2]. P2P Lending can be seen as an example of removal of intermediaries in financial section [3], [4]; as another technological interference for financial triggered by the existence of Internet; as a case of collaborative economics [5], or even as a platform to give lending to financially excluded people [6]. Because the traditional financial intermediary have been excluded, and dynamic environment that utilizes the collective intelligence of the crowd are made, P2P lending have the capability to reduce financial costs and improve financial market efficiency [2].

It is interesting to see how the credibility of the borrower can be filtered in this type of lending. Therefore many previous researches tried to predict the default of the lending data, from the Lending Club. There are some statistical techniques used to asses credit and predict the default. One of the most extensive technique that used by many researchers is logistic regression because it has the capability to predict with an accuracy that is not significantly different from the newer techniques [7]. The best results are that the test sample will be collected later than the training sample, to ensure inter-time validation. This has been done in this paper [1].

The aim of this study is to find what is the most significant attribute affecting default status based on Naïve Bayes and Decision Tree algorithm. Beside, we want to know which algorithm is the better with respect to the lending club data set. Using the data from Lending Club from 2017 until 2018, we found more than 1 million transaction data and filtered to 600 thousand data.

The structure of this paper is as follows. Section 2 gives the theoretical and empirical discussion related to P2P lending that has been done before, Section 3 concerns with data understanding and also the data modeling. Section 4 presents the results of the experiments of the default prediction, and finally in Section 5 the findings are discussed some conclusions were drawn from the whole process.

2. LITERATURE REVIEW

2.1 Naive Bayes

Naive Bayes is used to make a prediction to a certain variable based on another attribute [8] and it's subordinate of supervised learning based classification as did in [9,30,31]. This method has a dependent variable as the targeted attribute that influenced by the independent variable based on frequency and combination to generates future predictions [10]. In details the formula of bayes algorithm (figure 1) described as follows :

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

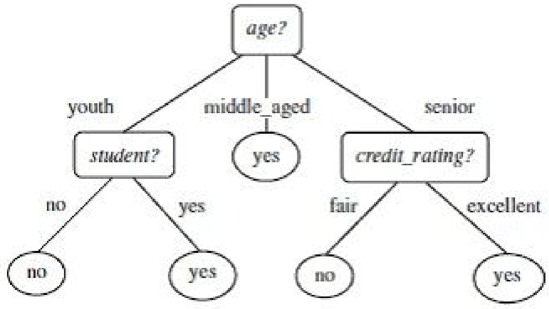
Figure 1: Naïve Bayes formula [10].

It also could be written as posterior = (prior x likelihood)/evidence [10]. Overall this method is one of common algorithm that has been widely used around the world due to its ability to accurately predict a target variable. Despite of its advantages, Naïve Bayes also has some potential drawback such as tend to have high bias results as pointed out in [11].

Naive Bayes algorithm is a probabilistic model that's used to produce prediction based on independent variables that will affect the targeted variable where it's have strong independence (bias) assumptions among the features.

2.2 Decision Tree

In general decision Tree is a data mining methodology that separates data to sub-set, and the purpose of this method is to filter, insert and process the data into sub-tree then find the group with the lowest cross-validation error [12]. This method furnishes a clear indication of what fields are most vital for forecast or classification & gives understandable rules [13].]. J48 is one of decision tree algorithm category where this method has advantages than ID3 such as manage data with missing value of variable, cutting decision tree after created, and it's still could reproduce normalized data with the smaller sub-tree [14]. To make it easier to understand, Figure 2 shows how decision tree J48 works in general



.

Figure 2: Concept of decision tree [14].

Decision tree is a classification technique that is commonly used used in machine learning to predict or classified data by separating each of it into sub-data and use the one with lowest error rate as the results. This method also have several variants such as ID3, Random Forest to mention a view. However this paper uses J48 algorithm to process the data since it's capability to normalized the data into smallest tree and have advantages to minimize pruning error.

2.3 CRISP-DM

This framework is common methodology that used as guidance in order to analyzed the business problems as well as the data [15]. First steps in this theory is to understand the business process to get the requirement for going into the data Understanding. Second, Data understanding is a process where to get essential data that match with the business goals. Third, Data preparation is to normalized or clean the data, derived attributes, and integrate the data. after normalized the data, the next step is to choose the algorithm that will give accurate results, these algorithms is depends of business & data needs then run the algorithm to get the results. Fifth, analyzed the results if the goals achieved then write the report but if not back to first step [16].

Based on explanation above, CRISP-DM have 5 steps ;business understanding, data understanding, data preparation, modeling, and evaluation to see if the data already match the expectation of business problems or not (Figure 3). As described before this method commonly used in the data mining world in order to start project and solve business problems with data driven approach.

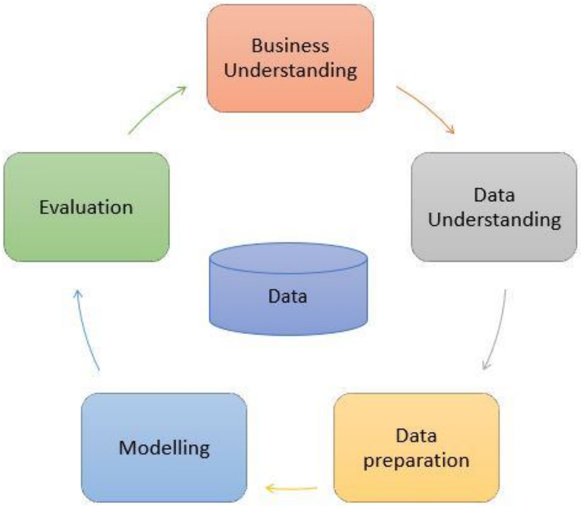


Figure 3: CRISP-DM Step.

2.4 P2P Lending

Peer-to-peer (P2P) lending are money lending activities from individuals to other individuals, without intermediaries from financial institutions [1]. Because of this, individuals allowed to lend and borrow directly on an Internet-based platform, without the involvement of financial intermediaries. P2P loans are risky activities because individuals lend their money to the other individual, not by companies which transfer credit risk. Credit risk is the potential of what impact any real or perceived changes financially in the credit worthiness of the borrower, while credit worthiness is the willingness and ability of the borrower to pay [17]. Credit scores are numbers that represent an assessment of a person's credit worth, or the likelihood that the person will pay his debt [18]. It is still confusing whether credit scoring will be another disruptige innovation,[19], but it is clear that P2P lending are fast spreading globally [20].

2.4 Related Works

Kumar et al [21], predict the default status of lending club with decision tree, random forest, and bagging. With 276169 datasets and 70 attributes. With the result in decision tree 81.3% accuracy, in random forest with 88.5%, and in bagging with 88.35% accuracy. Teply et al [22] with 212280 datasets and 23 attributes with the result on Naïve Bayes 78.36 % accuracy.

Yujin et al [23] predict 10 important feature through data-driven analysis with random forest and correlation matrix and classify the result into 3 categories.

Serrano cinca et al [1] analyze if the information provided by P2P Lending club which is a grade determined by them asymmetry with the reality happened. Resulting a positive result toward the hypothesis where the people with A Grade have 94.4% return rate and decreased gradually to 61.8% in G Grade. Xuchen Lin et al [24] developed a quantitative model of default prediction, with the attribute from previous research.

Qian et al [25] write a paper to predict default in P2P Lending with LSSVM and BABP Neural network with a result of 92.17% and 91.45%.

Li et al [26] proposed a multi-round model to predict default in P2P lending with XGBoost, DNN and LR for an organization in china. The AUC value result is 0.7869 and after some boosting and ensembling it increased to 0.7891.

Malekipirbazari and aksakalli [27] want to predict default in P2P lending with Random Forest, resulting a good accuracy but with a limitation from that approach is that this algorithm classified some of the good borrowers are bad.

Ge et al [28] study the relationship between social media information with the borrowers worthiness in online P2P Lending. Resulting two result. First, for all borrowers in P2P loans, the decision whether to disclose their social media accounts can be used as a predictor of their default probability. Second, for borrowers who choose to disclose their social media accounts, their social media presence, such as the scope of their social networks and the number of messages they post on social media sites, can predict the probability of default.

Fitzpatrick and Mues [29] compared four techniques for the purpose of predicting mortgage defaults. These two techniques are rooted in machine learning: Boosted Regression Trees (BRT) and Random Forests (RF), and the other two are statistical model: penalised Logistic Regression (LR) and semi-parametric Generalised Additive Models (GAMs). The result show that BRT and GAM are better in overall than the other algorithm

3. RESEARCH METHOD

For this research, the device used to run the prediction is PC with processor intel i7 6700 3.4 GHz, Nvidia RTX 2080 12 GB, 16GB memory, SSD 128GB, HDD 2TB.

3.1 Data Understanding

The data downloaded on the lending club's website where the date range from Q1 2016 - Q4 2018. This data contains transactions that occurred in Lending Club where attributes such as annual income, home_ownership, installment, grade, etc that will be useful as predictor variables in order to analyze probabilities of the future loan status.

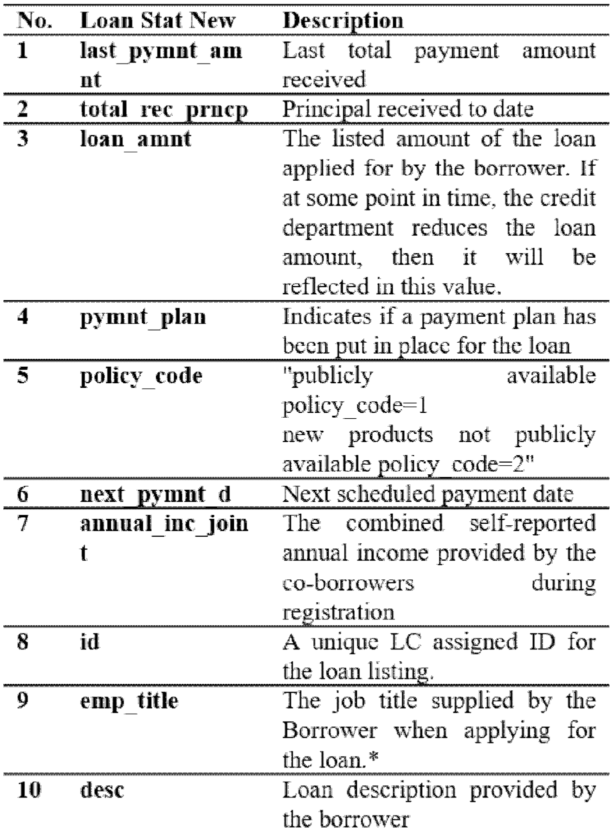
3.2 Data Preparation

Data combine from Q1 2016- Q4 2018 which contains 1 Million data. The data is cleaned, all data that contains "Current", "Late" as the loan_status value are deleted because those data cannot give information for default prediction, and data that contains "Charged Off" as loan_status values are changed to "Default".

After Data cleaning and filtering, the data remained to be used for more analysis, are 628408 data. After that we cleansed the data, by deletting some attribute. The attribute that contain only 1 data value will be deleted. The attribute that has blank data or zero data more than 90% of the total data, will be deleted. The most affecting attribute are job title, because it contains so many symbolic character that WEKA cannot read as attribute, and there are so many variance of data that will not affect the relation to P2P lending default prediction. The total attributes that we used after data cleansing and normalization is reduced from 152 to 86 attributes. Figure 4 presents the meaning of main 10 attributes out of 86 attributes that are used in this paper. A complete list of the attributes can be found in [32].

The attributes are then included in the analysis using several algorithms like Naive Bayes and J48 to see and compare the prediction accuracy, comparison of the confusion matrix, & ROC curve between those two algorithms.

Here some attributes that we delete because lack of data variation "pymnt_plan", "policy_code", "hardship_flag" ,etc. Data that a lot of blank "next_pymnt_d", "annual_inc_joint", "dti_joint",etc and some not important data such as "id","emp_title","url","desc",etc.



No.	Loan Stat New	Description
1	**last_pymnt_amnt**	Last total payment amount received
2	**total_rec_prncp**	Principal received to date
3	**loan_amnt**	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
4	**pymnt_plan**	Indicates if a payment plan has been put in place for the loan
5	**policy_code**	"publicly available policy_code=1 new products not publicly available policy_code=2"
6	**next_pymnt_d**	Next scheduled payment date
7	**annual_inc_joint**	The combined self-reported annual income provided by the co-borrowers during registration
8	**id**	A unique LC assigned ID for the loan listing.
9	**emp_title**	The job title supplied by the Borrower when applying for the loan.*
10	**desc**	Loan description provided by the borrower

Figure 4 Part of data dictionary of lending club attributes, listed only 10 out of 86 attributes used in this paper.

3.3 Data Modelling

In this research, the software that is used to do the modeling is Weka. The modeling techniques used for this research are Naive Bayes, Decision Tree.

The first method used is Naive Bayes because it assumes that the features are conditionally independent. Based on the training sample, the previous probability of each class and the conditional probability of each feature are obtained. This method has a dependent variable as the targeted attribute that influenced by the independent variable based on frequency and combination to generates future prediction .

The last one is the Decision tree, which is a powerful classification algorithm and has been widely used. There are many types of Decision Tree. In this research, we used the J48 model. decision tree is a classifier that used by machine learning to predict or classified data by separating each of it into sub-data and use the one with lowest error rate as the results. this method also have several category, however this research used J48 algorithm to process the data since it’s capability to normalized the data into smallest tree and have advantages to minimize pruning error. For this research, the configuration in J48 that used are minimum number of object 50, unpruned true, and useLaplace true.

4. DISCUSSION

To evaluate prediction performance, a 10-fold cross-validation approach is used, and the following metrics are obeserved; Prediction accuracy, Confusion matrix, and ROC Curve.

4.1 Naive Bayes

Using the Naive Bayes algorithm with loan_status as the targeted variable and 10 folds of cross-validation, the resulted confusion matrix is presented in Table 1. The accuracy of this test is 84.74%. Using Naïve Bayes we got that Area under ROC is 0.943. In figure 5 the X axis gives the data about false positive rate and Y axis represents the data about true positive rate. In figure 6 the X axis give the data about instance number rate and Y axis give the data about precision. This algorithm ran for 8.29 seconds to build the model.

Table 1: Confusion Matrix Naïve Bayes

	Predicted Default	Predicted Fully Paid
Condition Default	123622	11755
Condition Fully Paid	84096	408935

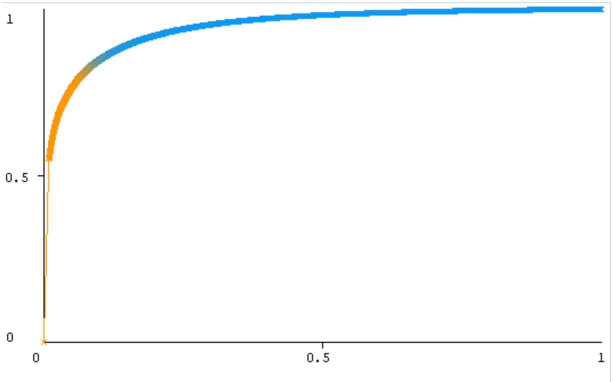


Figure 5: False positive rate (x) compares to True positive rate (y) of Naïve Bayes algorithm. (ROC Curve)

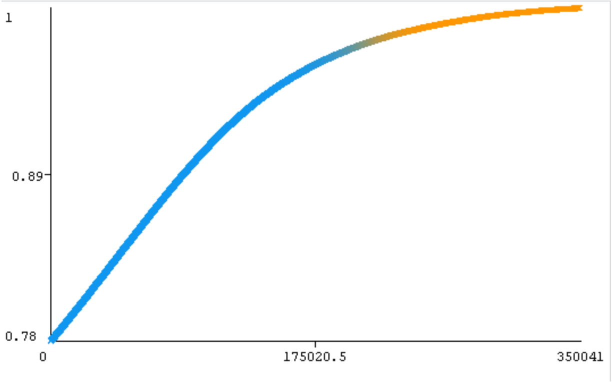


Figure 6: instance number rate (x) compares to precision (y) of Naïve Bayes algortihm.

4.2 Decision Tree

The decision tree algorithm that we use is J48 with some parameter changing, we change minimum number of object from 2 to 50 and we also change unpruned and useLaplace from false to true for Peer to Peer lending dataset. The model of J48 algorithm shows that "last_pymnt_amnt" attribute as the primary split then continues with "total_rec_prncp", "loan_amnt", etc. The result of this test is 99,88 %. Based on the test result, the most influence variables are total_rec_prncp, last_payment, loan_amnt, installment, and funded_amnt_inv. It took 184.35 seconds to build the model. From that model we get number of leave is : 1341 and size of the tree is 1433.

For figure 7, X axis represnts the data about false positive rate and Y axis gives the data about true positive rate resulting from applying J48 to the dataset. In Figure 8 X axis gives the data about instance number rate and Y axis gives the data about precision.

Table 2: Confusion matrix decision tree

	Predicted Default	Predicted Fully Paid
Condition Default	134403	974
Condition Fully Paid	103	492928

The confusion matrix obtained from implementing the J48 algorithm to the dataset is shown in Table 2. It can be seen that most of the samples fall in the main diagonal.

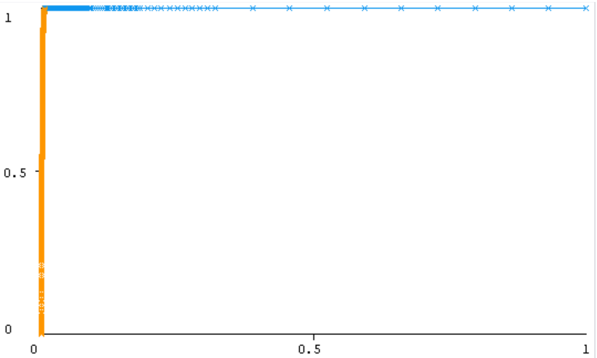


Figure 7: False positive rate (x) compares to True positive rate (y) of J48 Algorithm. (ROC)

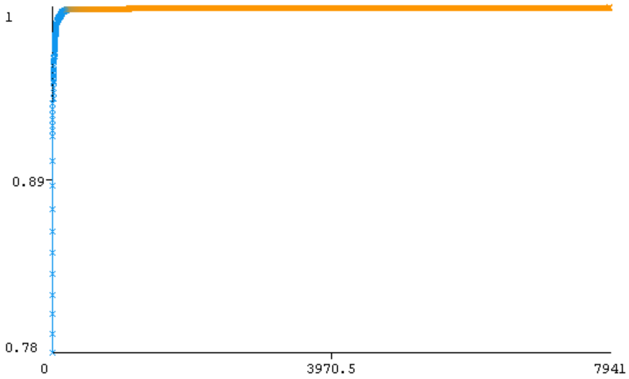


Figure 8: instance number rate (x) compares to precision (y) of J48 Algorithm.

From this research we could see that from this two algortihm, decision tree resulting in better accuracy and AUC area. But it takes longer time to run.

5. CONCLUSION

From this research, financial technology like online peer-to-peer (P2P) lending is the current trend that already disturbed the finance industry. However, this industry has a great risk that needs to be managed in purpose to survive in the market. This research use data from landing club to predict & mitigate default risk of online P2P lending.

From the experiments above, decision tree J48 has performed very well to predict the default risk, where the accuracy is 99%, an average of ROC is 0.99, and the most influencing attributes are Total_rec_prncp, last_payment, loan_amnt, installment, and funded_amnt_inv. Next is Naive Bayes that has 84.74% accuracy we got that Area under ROC is 0.943. It is absolutely better than previous research, but there are some weakeness of this research, the tendencies of overfitting are very high, because of the data cleaning that is not very good. However the accuracy is better then the previous research

For future research, it's interesting to see which factors that have the influence to accept future lending from new users because it could prevent larger default risk on the entry phase, looking to solve the tendencies of overfitting.

REFERENCES

1. C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "**Determinants of default in P2P lending**," *PLoS One*, vol. 10, no. 10, pp. 1–22, 2015. https://doi.org/10.1371/journal.pone.0139427
2. Y. Guo, W. Zhou, C. Luo, C. Liu, and H. Xiong, "**Instance-based credit risk assessment for investment decisions in P2P lending**," *Eur. J. Oper. Res.*, vol. 249, no. 2, pp. 417–426, 2016. https://doi.org/10.1016/j.ejor.2015.05.050
3. D. Fiaschi, I. Kondor, M. Marsili, and V. Volpati,

"**The interrupted power law and the size of shadow banking**," *PLoS One*, vol. 9, no. 4, pp. 1–8, 2014. https://doi.org/10.1371/journal.pone.0094237

4. E. Lee and B. Lee, "**Herding behavior in online P2P lending: An empirical investigation**," *Electron. Commer. Res. Appl.*, vol. 11, no. 5, pp. 495–503, 2012 https://doi.org/10.1016/j.elerap.2012.02.001.
5. R. Belk, "**You are what you can access: Sharing and collaborative consumption online**," *J. Bus. Res.*, vol. 67, no. 8, pp. 1595–1600, 2014. https://doi.org/10.1016/j.jbusres.2013.10.001
6. H. Yum, B. Lee, and M. Chae, "**From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms**," *Electron. Commer. Res. Appl.*, vol. 11, no. 5, pp. 469–483, 2012. https://doi.org/10.1016/j.elerap.2012.05.003
7. L. C. Thomas, "**Consumer finance: Challenges for operational research**," *J. Oper. Res. Soc.*, vol. 61, no. 1, pp. 41–52, 2010.
8. M. M. Saritas, "**Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification**," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019. https://doi.org/10.18201/ijisae.2019252786
9. J. August, I. Journal, J.- August, and P. Mittal, "**Data Mining Techniques for IoT enabled Smart Parking Environment : Survey**," International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 4, 2019.
10. Y. C. Zhang and L. Sakhanenko, "T**he naive Bayes classifier for functional data**," *Stat. Probab. Lett.*, vol. 152, pp. 137–146, 2019. https://doi.org/10.1016/j.spl.2019.04.017
11. L. P. Maguluri and L. Ragupathy, "**A New sentiment score based improved Bayesian networks for real-time intraday stock trend classification**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 4, no. 1, pp. 1–9, 2019.
12. Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "**A decision tree method for building energy demand modeling**," *Energy Build.*, vol. 42, no. 10, pp. 1637–1646, 2010.
13. H. Hong *et al.*, "**Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area** (China)," *Catena*, vol. 163, no. July 2017, pp. 399–413, 2018.
14. R. Panigrahi and S. Borah, "**Rank Allocation to J48 Group of Decision Tree Classifiers using Binary and Multiclass Intrusion Detection Datasets**," *Procedia Comput. Sci.*, vol. 132, pp. 323–332, 2018.
15. S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "**DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model**," *Procedia CIRP*, vol. 79, pp. 403–408, 2019. https://doi.org/10.1016/j.procir.2019.02.106
16. R. Wirth, "**CRISP-DM : Towards a Standard Process Model for Data Mining**," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.
17. M. Crouchy, D. Galai, and M. Robert, ***The essentials of risk management,*** Vol 1. New York: McGraw-Hill, 2006.
18. S. Mehmood, "[WIP] Mp r a," ***Econ. Policy,*** no. 2116, pp. 0–33, 2013.
19. C. M. Christensen and M. Overdorf, "**Meeting the challenge of disruptive change**," *Harv. Bus. Rev.*, vol. 78, no. 2, 2000.
20. G. Ruiqiong and F. Junwen, "**An Overview Study on P2P Lending**," *Int. Bus. Manag.*, vol. 8, no. 2, pp. 14–18, 2014.
21. V. K. L, S. Natarajan, S. Keerthana, K. M. Chinmayi, and N. Lakshmi, "**Credit Risk Analysis in Peer-to-Peer Lending System** Vinod Kumar L Keerthana S , Chinmayi K M , Lakshmi," *2016 IEEE Int. Conf. Knowl. Eng. Appl.*, pp. 193–196, 2016.
22. P. Teply and M. Polena, "**Best classification algorithms in peer-to-peer lending**," *North Am. J. Econ. Financ.*, no. January, 2019. https://doi.org/10.1016/j.najef.2019.01.001
23. Y. Jin and Y. Zhu, "**A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending**," *Proc. - 2015 5th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2015*, pp. 609–613, 2015.
24. X. Lin, X. Li, and Z. Zheng, "**Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China**," *Appl. Econ.*, vol. 49, no. 35, pp. 3538–3545, 2017. https://doi.org/10.1080/00036846.2016.1262526
25. M. Qian and F. Hu, "**An Empirical Study on Prediction of the Default Risk on P2P Lending Platform**," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 490, no. 6.
26. W. Li, S. Ding, Y. Chen, and S. Yang, "**Heterogeneous ensemble for default prediction of peer-to-peer lending in China**," *IEEE Access*, vol. 6, no. c, pp. 54396–54406, 2018.
27. M. Malekipirbazari and V. Aksakalli, "**Risk assessment in social lending via random forests**," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, 2015. https://doi.org/10.1016/j.eswa.2015.02.001
28. R. Ge, J. Feng, B. Gu, and P. Zhang, "**Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending**," *J. Manag. Inf. Syst.*, vol. 34, no. 2, pp. 401–424, 2017. https://doi.org/10.1080/07421222.2017.1334472
29. T. Fitzpatrick and C. Mues, "**An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market**," in *European Journal of Operational Research*, 2016, vol. 249, no. 2, pp. 427–439. https://doi.org/10.1016/j.ejor.2015.09.014

30 Gunawan, F. E., Soewito, B., Mauritsius, T., & Surantha, N. (2018, December). **Vibration-based classification of road damages: gyroscope data and**

a simple neural network model. In IOP Conference Series: Earth and Environmental Science (Vol. 195, No. 1, p. 012068). IOP Publishing. https://doi.org/10.1088/1755-1315/195/1/012068

31. Radiansyah, F., & Mauritsius, T. (2019). **Condition-based maintenance using data mining techniques on internet of things generated data**. Journal of Theoretical and Applied Information Technology, 97(13), 3702-3717.
32. Lending Club (2019). **Lending Club Data Dictionary [online].** Resources.lendingclub.com. Available at: https://resources.lendingclub.com/LCDataDictionary.xlsx [Accessed 13 Sep. 2019].