Original Article

Prophecy of loan approval by comparing Decision Tree with Logistic Regression, Random Forest, KNN for better Accuracy.

B. Aditya[1], V. Nagaraju[2]

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.
[2]Project Guide, Corresponding Author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

Abstract

Aim: The aim of the work is to evaluate the accuracy and cross validation in predicting loan approval using Novel Penalty Based Approach Logistic Regression (LR), Random Forest (RF), Decision Tree (DT) and K-Nearest Neighbor (KNN) Classification algorithms. **Materials and Methods**: The classification algorithm is invoked on a loan approval dataset consisting of 615 records. A framework for forecasting loan approval in the banking sector has been proposed and developed that compares Novel Penalty Based Approach Logistic Regression, decision tree, Random Forest and K-Nearest Neighbor classifiers. Sample size was calculated as 55 in each group using G powers. Sample size was calculated using clinical analysis, with alpha and beta values of 0, 05 and 0. 5, 95% confidence, 80% pre-test g power and enrolment ratio 1. **Results**: The Novel Penalty Based Approach Logistic Regression classifier produces 77.27% accuracy in predicting the Loan Approval on the data set, whereas the Random Forest, K-Nearest Neighbor, Decision Tree classifiers produce 76.62%, 72.27%, 72.07% respectively. The significant value is 0.016. Hence Novel Penalty Based Approach Logistic Regression is better than KNN, RF, DT classifiers. **Conclusion:** The results show that the performance of Novel Penalty Based Approach Logistic Regression is better when compared with KNN, RF and DT in terms of both cross validation and accuracy.

Keywords: Classification, Novel Penalty Based Approach, Logistic Regression, Decision Tree, K-Nearest Neighbor, Random Forest, Loan Approval, Machine Learning, Data Mining.

DOI: 10.47750/pnr.2022.13.S03.87

INTRODUCTION

Loan distribution is the main part of the business of almost all banks. Most of the bank's assets come directly from profits from loans granted by banks. The main goal in the banking environment is to invest your assets in safe hands where they are (Bhagat 2018). Today many banks / financial companies approve the loan after a process of verification and validation of appeals but there is still no certainty that the chosen applicant is the right deserving applicant among all applicants. Using this system, we can predict whether that particular requestor is safe or not and the whole feature validation process is automated by machine learning technique. The downside of this model is that it emphasizes different weights for each factor, but in real life sometimes the loan can only be approved on the basis of one strong factor (Xu, Lu, and Xie 2021), which is not is not possible with this system. The loan forecast is very useful for the bank employees as well as for the applicant. The purpose of this document is to provide a quick, immediate and easy way to select deserving candidates. It can offer special advantages to the bank. The loan forecasting system can automatically calculate the weight of each characteristic that participates in loan processing, and the same characteristics are processed on the new test data against their associated weight (Chen, Zhang, and Ng 2018). It is possible to set a deadline for the applicant to check whether his loan can be sanctioned or not. The loan forecasting system allows you to jump to a specific question so that it can be checked first. This document is exclusively intended for the managing authority of the bank / financial company, the entire forecasting process is done in private, no interested party can modify the treatment (Karthiban, Ambika, and Kannammal 2019). The result of a particular loan number or loan ID can be sent to various banking services so that they can take appropriate action upon request. The proposed work is helpful for managers of the particular banks to reduce the time taking processes for approving loans.

Approximately 53 related articles published in IEEE Xplore and 28 related articles have been published related to this work in Google Scholar. The scoring algorithm is widely used to improve the predictability of loan approval in a variety of banking and financial industries. In (Sheikh, Goel, and Kumar 2020), the performance of the system is measured in terms of classification accuracy and the results reveal that it has great potential to accurately predict the status of loan approval. (Ye, Dong, and Ma 2018) proposed a feature selection algorithm with a classifier K-Nearest Neighbor to design a high-level intelligent system to predict loan approval. (Stahringer 2012) proposed an algorithm for the accuracy level of loan approval predictions of 80% by Novel Penalty Based Approach Logistic Regression. (Kemalbay and Korkmazoğlu 2014) proposed spatial clustering based on application criteria with noise and decision tree executed by other models reaching 79% to predict loan approval. Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021)

Several works have shown that KNN, RF and DT performance is poor and offers less accuracy in predicting loan approval. A study compares the accuracy of various data mining classification algorithms in loan approval prediction. It is important to analyze and compare various classification algorithms that provide better cross validation.The research gap in the existing method is that the Decision Tree has low accuracy for predicting loan approval. Hence the proposed method aims at comparing algorithms to know which algorithm was giving more accuracy than the Decision Tree. Therefore, the work aims to compare the accuracy of LR, RF, DT and KNN algorithms in forecasting loan approval.

Materials And Methods

The research work was carried out at Web Ontologies Lab, Department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS. Work was performed on 615 records from a loan dataset. The accuracy of loan approval forecasts was obtained by evaluating four groups. A total of 10 iterations were performed on each group to achieve better accuracy. The dataset was downloaded from the Kaggle website. The dataset contains 615 rows and 13 columns. Some of the important attributes taken for the experimental setup are loan ID, loan amount, employment status, etc. (navaneethkarumuru 2021) .

The sample size was calculated to be 22 in each group using the G power. In (Rembart and Soliman 2017) the loan approval dataset was used with a sample size of 615 customers, 13 characteristics and some missing values. Sample size was calculated using loan amount analysis, with alpha and beta values of 0. 05 and 0. 5, 95% confidence, pre-test power of 80% and an enrollment rate of 1.

Novel Penalty Based Approach Logistic Regression (LR)-Group 1

Input: Loan Approval dataset

Output: Accuracy

Step 1. Import and read the dataset
Step 2. Select the features randomly from the dataset
Step 3. Generate the LR classifier penalty as a parameter.
Step 4. l2 was used as a parameter value.
Step 5. Analyze the dataset by varying dependent and independent variables
Step 6. LR predicts the outcome in a categorical variable.
Step 7. Finally predicts the possibility of event using the log function.

This study uses the Novel Penalty Based Approach Logistic Regression class from sklearn. linear model library. This innovative method uses a penalty-based approach. It takes the penalty as a parameter. "LR" is used as the parameter value. The data set is randomly divided into training (80%) and test (20%). Select samples at random and analyze the data set by varying the dependent variables and independent variables. Finally, it provides for the possibility of an event using the log function (Vaidya 2017).

Decision Tree (DT) - Group 2

Input: Loan Approval dataset

Output: Accuracy

Step 1. Import and read the dataset
Step 2. Select the features randomly from the dataset
Step 3. Generate the DT classifier criteria as a parameter.
Step 4. Gini was used as a parameter value.
Step 5. Construct a decision tree using DT classifier and predict the result for every sample.

Step 6. Voting was performed for every predicted result.
Step 7. Most predicted results were selected as final output.

This study uses the Decision Tree Classifier class from the sklearn library. It takes citian as a parameter. "Gini" is used as the parameter value. The data set is randomly divided into training (80%) and test (20%). It selects the samples at random and decision trees were collected for each sample to predict the outcome. Voting was taken for each predicted outcome and the most voted outcome was selected as the final outcome. The work implements an innovative method that uses a criteria-based decision tree classifier (Thu 2020).

Random Forest Classifier (RF) - Group 3

Input: Loan Approval dataset
Output: Accuracy
Step 1. Import and read the dataset
Step 2. Select the features randomly from the dataset
Step 3. Generate the RF classifier criterion as a parameter.
Step 4. Gini was used as a parameter value.
Step 5. Construct a DT using RF classifiers and predict the result for every sample.
Step 6. Voting was performed for every predicted result.
Step 7. Most voted prediction results were selected as the final outcome.

In this study, the Random Classifier class of the sklearn ensemble library is used. It takes citerian as a parameter. “Gini” is used as the parameter value. The dataset is splitted randomly into training (80%) and testing (20%). It selects samples randomly and the decision trees were collected for every sample to predict the result. Voting was performed for every predicted result and the most voted result was selected as the final result. The algorithm uses a Random Forest Classifier (NTSRF) (Madaan et al. 2021).

K-Nearest Neighbor(KNN) - Group 4

Inputs: Loan Approval data set
Output: Accuracy.
Step 1. Load the dataset
Step 2. Split the dataset randomly into training (80%) and testing (20%) dataset
Step 3. Set the target variable
Step 4. Generate the KNN classifier based on the training set
Step 5. Train the classifier using rbf kernel parameter
Step 6. Predict the testing set based on training dataset
Step 7. Evaluate the classifier.
Step 8. Return Accuracy.

K-Nearest Neighbor(KNN) is a regulated machine learning algorithm which can be utilized for both classification and regression challenges. In this study, to train the KNN the neighbor class of scikit learn library was used. Import the loan approval. csv dataset and load the dataset. The dataset is split randomly into training (80%) and testing (20%) sets. The target variable is selected. Then, the KNN classifier based on the training set is generated. Rbf was used as the value of the kernel parameter. The testing set is predicted based on the training set. The KNN classifier is evaluated and the accuracy is calculated.

The proposed work was experimented in Jupyter, The Hardware and Software requirements for experimenting the work includes i3 processor, 50 GB HDD, 4 GB RAM, Windows OS, Python: Colab/Jupyter. Initially, the data set was divided into two parts: the training and test sets. Then, the algorithm is tested on the training and test sets. The training and testing sets are changed 10 times depending on the size of the test set. Table 1 shows the comparison between the accuracy and cross validations of KNN and LR for 10 iterations. The different parameters for the analysis can be calculated as follows:

Accuracy :- It identifies the number of instances that were correctly classified as shown in the following equation 1.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (1)$$

Cross Validation is used to calculate which part of prediction data is positive using equation 2.

$$\text{Cross Validation} = \frac{TP}{TP+FP} \quad (2)$$

Statistical Analysis:

In addition to the experimental analysis, the work was statistically evaluated using the Statistical Package for the Social Sciences (SPSS). Analysis was performed to obtain the mean, standard deviation and mean of standard errors. To compare the parameters of the four groups, a T test of the independent variable was performed. In the analysis, the independent variables are marital status, loan identifier, address, sex. The dependent variables that affect the output are cross validation and cross-validation (Abakarim, Lahby, and Attioui 2018). Fig. 5 shows the accuracy graph between LR, RF, DT and KNN.

Results

Table 1 shows the comparison between the accuracy and cross validation of four groups for 10 iterations. Fig. 1 shows the graph of total income log. Table 2 shows the different parameters of the four groups. Accuracy and cross-validation were calculated for KNN, DT, RF and LR. Four-group analysis shows that LR has higher cross validation (77.27%) and cross-validation (80.62%) than RF, DT, KNN.

Table 3 shows the statistical analysis of LR, DT, RF and KNN with different sets of test data. Fig. 2 shows the graph of loan amounts taken by the applicants. Fig. 3 shows the graph of applicant income for loan approval status and Fig. 4 shows the loan approval status. The average accuracy of the LR model appears to be higher than that of the RF, DT and KNN models. In addition, the cross-validation of LR is much better than that of RF, DT and KNN. The performance of the LR algorithm is superior to that of the LR, DT and KNN algorithm. There are no significant differences between the four groups. Therefore, LR is better than DT, RF and KNN.

Table 2 illustrates the mean accuracy and cross validation of KNN, RF, DT and LR. Statistical analysis of four independent groups shows that LR has a higher mean accuracy (77. 27%) and mean cross validations (80. 26%) than KNN. The average error of KNN, RF, DT is slightly less than LR.

Discussion

The work demonstrates that Novel Penalty Based Approach Logistic Regression is better than KNN, RF, DT at predicting loan approval in terms of accuracy and cross validation. However, the average error of LR appears to be higher than KNN. and the banking sectors. Experimental work was carried out between four groups KNN, RF, DT and LR by varying the size of the test. Based on the experimental results performed in the Jupyter Notebook, the accuracy and cross-validation of LR is 77.27% and 80.62%, while KNN provides the accuracy and cross-validation of 67.2%. This shows that LR is better than KNN, RF and DT. Various parameters such as cross validation are also compared. From the SPSS plot, the proposed Novel Penalty Based Approach Logistic Regression classifier offers better performance in terms of cross validation (77.27%) and cross-validation (80.62%) compared to KNN, RF and DT algorithms. Fig. 5. shows that the average error of LR turns out to be slightly greater than KNN, RF and DT which must be minimized.

The most important aspect of loan approval forecasting is accuracy and cross-validation. The (Ambika, Ambika, and Biradar 2021) study proposed a machine learning-based classification system for predicting loan approval based on a loan approval record. The (Isik, Tastan, and Yolum 2007) study used selection algorithms, the cross-validation method, and seven performance scoring metrics for classifiers, such as classification cross validation, specificity, sensitivity, Matthews correlation coefficient, and execution time. suggested to predict the loan approval. (Gopu 2021) The Random Forest algorithm was used in the Spark framework to predict loan approval and showed that the study achieved a higher accuracy rate even with a dataset of 613 documents. In the study by (Mortensen et al. 1988) , attribute filtering, frequent item mining and a variety of data mining techniques such as Decision tree, Naive Bayes, Support vector machine and SVM classifications are used for predicting loan approval status. When it comes to predicting loan approval, the accuracy of Naive Bayse was superior to that of other Data mining algorithms. The accuracy of the LR classification algorithm depends on the size of the training and testing data set. In our study, the precision and cross-validation appear to be better than KNN. However, the average error appears to be higher in our proposed work which should be minimized.

Although the results of the study are better in both experimental and statistical analysis, there are some limitations in the work. Accuracy assessment cannot provide a better result on larger data sets. Still in LR, the average error seems to be greater than KNN. It would be preferable if the average error could be reduced considerably. However, the work can be improved by applying optimization algorithm techniques, to achieve better precision and less average error as a future work. Feature selection algorithms can be used prior to classification to improve the classification accuracy of classifiers. Therefore, thanks to the Data Mining algorithms, we can reduce the computation time and improve the accuracy of the classification of classifiers.

Conclusion

Novel Penalty Based Approach Logistic Regression Classifier is a classification technique that uses averaging to improve the accuracy and cross validation. The work shows that the Accuracy and Cross Validation for loan approval prediction using Novel Penalty Based Approach Logistic Regression (LR) is better than the K-Nearest Neighbor (KNN), Random Forest (RF) and Decision Tree (DT). It is found that LR performs significantly better than KNN, RF and DT in predicting loan approval accurately, but the mean error is found to be little higher than KNN, RF and DT. Hence, it is concluded that LR classifier results in acceptable accuracy and cross validation than KNN, RF and DT.

DECLARATIONS

Conflicts of interest

No conflict of interest in this manuscript.

Authors Contribution

Author BJSA was involved in data collection, data analysis, algorithm framing, implementation and manuscript writing. Author VN was involved in designing the work flow, guidance and review of manuscript.

Acknowledgements

We thank Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (formerly Saveetha University) for providing facilities and continued assistance to complete this study.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. ZONIAC Software Pvt. Ltd.
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering.

References

1. Abakarim, Youness, Mohamed Lahby, and Abdelbaki Attioui. 2018. "Towards An Efficient Real-Time Approach To Loan Credit Approval Using Deep Learning." 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC). https://doi.org/10.1109/isivc.2018.8709173.
2. Adhinarayanan, Rajesh, Aravindh Ramakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and Damodharan Dillikannan. 2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, June, 1–18.
3. Ambika, Ambika, and Santosh Biradar. 2021. "Survey on Prediction of Loan Approval Using Machine Learning Techniques." International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-1165.
4. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." Biomass Conversion and Biorefinery, August. https://doi.org/10.1007/s13399-020-00938-0.
5. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique." Biomass Conversion and Biorefinery, May. https://doi.org/10.1007/s13399-021-01551-5.
6. Bhagat, Abhishek. 2018. Predicting Loan Defaults Using Machine Learning Techniques.
7. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." Fuel 304 (November): 121490.
8. Chen, Ya-Qi, Jianjun Zhang, and Wing W. Y. Ng. 2018. "Loan Default Prediction Using Diversified Sensitivity Undersampling." 2018 International Conference on Machine Learning and Cybernetics (ICMLC). https://doi.org/10.1109/icmlc.2018.8526936.
9. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as

Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." Biomass Conversion and Biorefinery, February. https://doi.org/10.1007/s13399-021-01312-4.

10. Gopu, A. P. 2021. "Data Analysis of Customer Complaints of Loan Approval and Financing in Banking Industry Using SVM Classification." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.33520.
11. Isik, Feyza Merve, Bulent Tastan, and Pinar Yolum. 2007. "Automatic Adaptation of BPEL Processes Using Semantic Rules: Design and Development of a Loan Approval System." 2007 IEEE 23rd International Conference on Data Engineering Workshop. https://doi.org/10.1109/icdew.2007.4401089.
12. Jayanth, Bellappu Venkat, Melvin Victor Depoures, Gopal Kaliyaperumal, Damodharan Dillikannan, Dilipsingh Jawahar, Kumaran Palani, and Ganesha Prasad Meravanigee Shivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, June, 1–18.
13. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." Process Biochemistry 99 (December): 36–47.
14. Karthiban, R., M. Ambika, and K. E. Kannammal. 2019. "A Review on Machine Learning Classification Technique for Bank Loan Approval." 2019 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/iccci.2019.8822014.
15. Kemalbay, Gülder, and Özlem Berak Korkmazoğlu. 2014. "Categorical Principal Component Logistic Regression: A Case Study for Housing Loan Approval." Procedia - Social and Behavioral Sciences. https://doi.org/10.1016/j.sbspro.2013.12.537.
16. Madaan, Mehul, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. 2021. "Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study." IOP Conference Series: Materials Science and Engineering. https://doi.org/10.1088/1757-899x/1022/1/012042.
17. Mortensen, Timothy L., David L. Watt, F. Larry Leistritz, and North Dakota Agricultural Experiment (Fargo). 1988. Loan Default Prediction Using Logistic Regression and a Loan Pricing Model.
18. navaneethkarumuru. 2021. "LoanPrediction." Kaggle. April 10, 2021. https://kaggle.com/navaneethkarumuru/loanprediction.
19. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." Fuel 277 (October): 118166.
20. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." Fuel 278 (October): 118315.
21. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." Biomass Conversion and Biorefinery, March. https://doi.org/10.1007/s13399-021-01427-8.
22. Rembart, Franz, and Erfan Soliman. 2017. "Loan Demand in Jordanian Microfinance Market: Interest Rate Elasticity and Loan-Acceptance Prediction via Logistic Regression." Enterprise Development and Microfinance. https://doi.org/10.3362/1755-1986.16-00009.
23. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." Environmental Progress & Sustainable Energy 40 (6). https://doi.org/10.1002/ep.13696.
24. Shanmugam, Rajasekaran, Damodharan Dillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 43 (23): 3064–81.
25. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. 2020. "An Approach for Prediction of Loan Approval Using Machine Learning Algorithm." 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). https://doi.org/10.1109/icesc48915.2020.9155614.
26. Stahringer, Tim. 2012. Logistic Regression Applied on Loan Default Prediction Using Peer-to-Peer Lending Data.
27. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of Ganoderma Lucidum Using Saccharomyces Cerevisiae." Fuel 306 (December): 121680.

28. Thu, Thuy Nguyen Thi. 2020. "Machine Learning Solution in Peer to Peer Loan Service in Vietnamese Banks." Journal of Advanced Research in Dynamical and Control Systems. https://doi.org/10.5373/jardcs/v12sp7/20202268.
29. Vaidya, Ashlesha. 2017. "Predictive and Probabilistic Approach Using Logistic Regression: Application to Prediction of Loan Approval." 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt.2017.8203946.
30. Xu, Junhui, Zekai Lu, and Ying Xie. 2021. "Loan Default Prediction of Chinese P2P Market: A Machine Learning Methodology." Scientific Reports. https://doi.org/10.1038/s41598-021-98361-6.
31. Ye, Xin, Lu-An Dong, and Da Ma. 2018. "Loan Evaluation in P2P Lending Based on Random Forest Optimized by Genetic Algorithm with Profit Score." Electronic Commerce Research and Applications. https://doi.org/10.1016/j.elerap.2018.10.004.**Tables and Figures**

Table 1. Accuracy and Cross Validation achieved during evaluation of Loan Approval prediction using test dataset with, KNN algorithm, DT, RF and LR technique for different iterations.

Iteration No.	**ACCURACY**				**CROSS VALIDATION**			
	LR	**DT**	**RF**	**KNN**	**LR**	**DT**	**RF**	**KNN**
1.	77. 27	70. 77	77. 27	72. 27	80. 62	70. 03	79. 31	75. 89
2.	75. 33	60. 66	73. 33	73. 33	80. 46	70. 79	79. 80	76. 80
3.	80. 00	80. 00	74. 90	74. 90	82. 88	80. 77	78. 60	76. 60
4.	80. 00	73. 33	75. 80	72. 80	79. 94	74. 26	79. 09	77. 09
5.	80. 01	75. 24	76. 98	73. 98	78. 42	70. 02	79. 98	78. 98
6.	77. 27	70. 12	76. 99	74. 99	80. 62	71. 01	78. 99	77. 99
7.	78. 93	71. 00	75. 98	71. 98	79. 45	74. 10	76. 98	75. 98
8.	77. 77	72. 37	76. 88	73. 88	80. 01	75. 40	79. 79	76. 79
9.	77. 01	72. 01	74. 00	71. 00	79. 87	74. 99	77. 78	74. 78
10.	78. 00	73. 00	74. 92	71. 92	80. 61	74. 80	78. 02	75. 02

Table 2. Experimental analysis in Jupyter for Accuracy, Cross Validation for DT, RF, KNN and LR. LR provides better Accuracy (77. 27%) and cross validation (80. 62%) than KNN, DT and RF.

MODEL	**ACCURACY**	**CROSS VALIDATION**
LR	77.27	80.62
DT	72.07	70.04
RF	76.62	79.97
KNN	72.27	75.89

Table 3. Statistical Analysis of Mean, Standard Deviation and Standard Error of cross validation and Accuracy of LR and KNN algorithms. There is a statistically significant difference in Cross Validation and accuracy values between the data mining algorithms. LR had the higher Cross Validation (80. 62%) and accuracy (77. 27%) and KNN had Cross Validation(80. 46%) and accuracy (76. 43%)

	Algorithm				**Std.**	**Std. Error**	**95% Confidence Interval for Mean**

		N	Mean	sig	Deviation	Mean	Lower	Upper
ACCURACY	LR KNN DT RF total	10 10 10 10 40	77. 5990 73. 1050 71. 8500 75. 7050	. 016	1. 42165 1. 33680 4. 85136 1. 37450	. 44957 . 42273 1. 53414 . 43466	3. 25752 3. 25717 3. 75587 3. 51592	5. 85048 5. 85083 5. 86213 5. 27208
CROSS VALIDATION	LR KNN DT RF total	10 10 10 10 40	80. 1630 76. 5920 73. 6170 78. 8340	. 044	1. 34895 1. 27407 3. 31443 0. 98559	. 42657 . 40290 1. 04812 . 31167	1. 96325 1. 96297 1. 96257 1. 96342	4. 42875 4. 42903 4. 42764 4. 42536

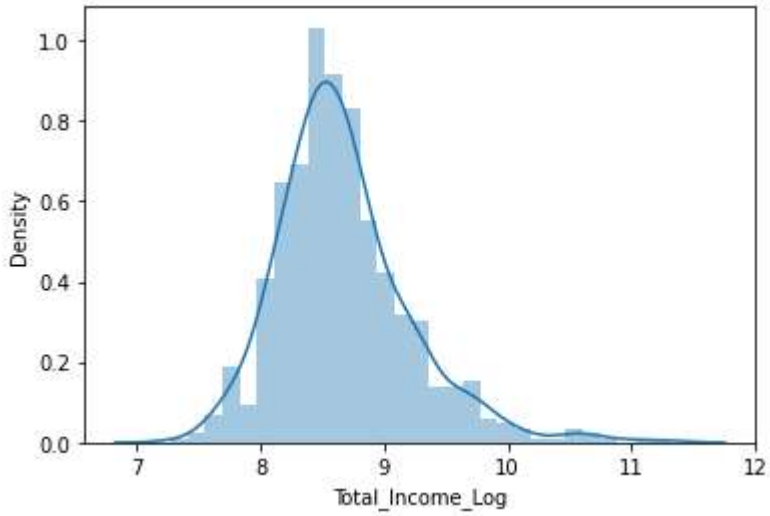

Fig. 1. Total Income log for Loan Prediction

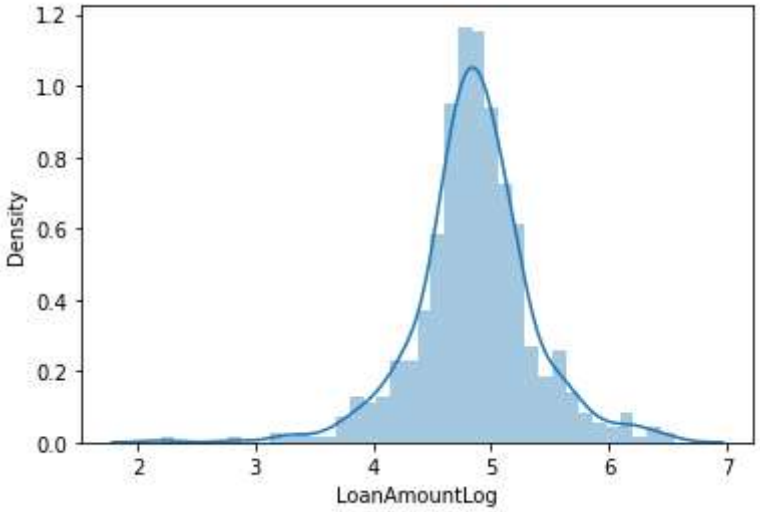

Fig. 2. Loan Amount Log for Loan Approval Status

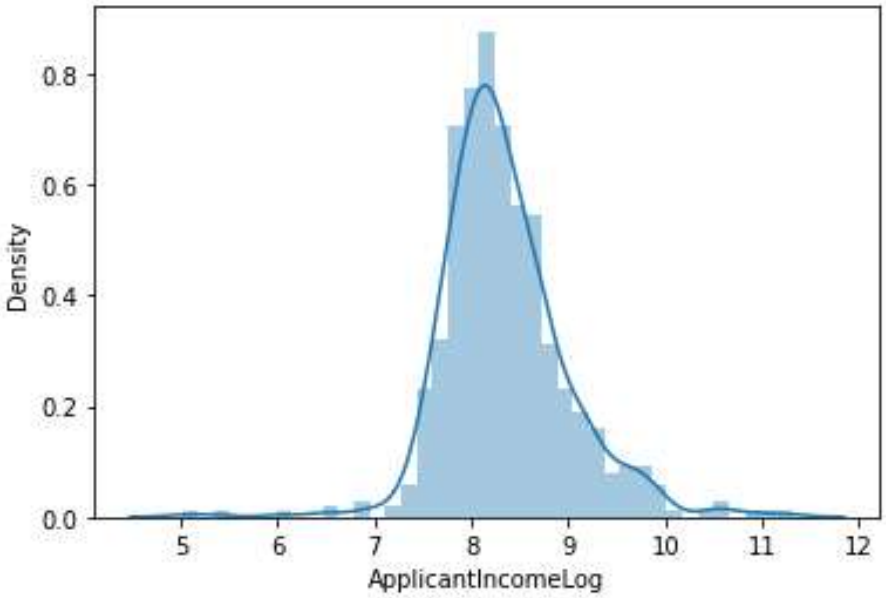

Fig. 3. Applicant Income

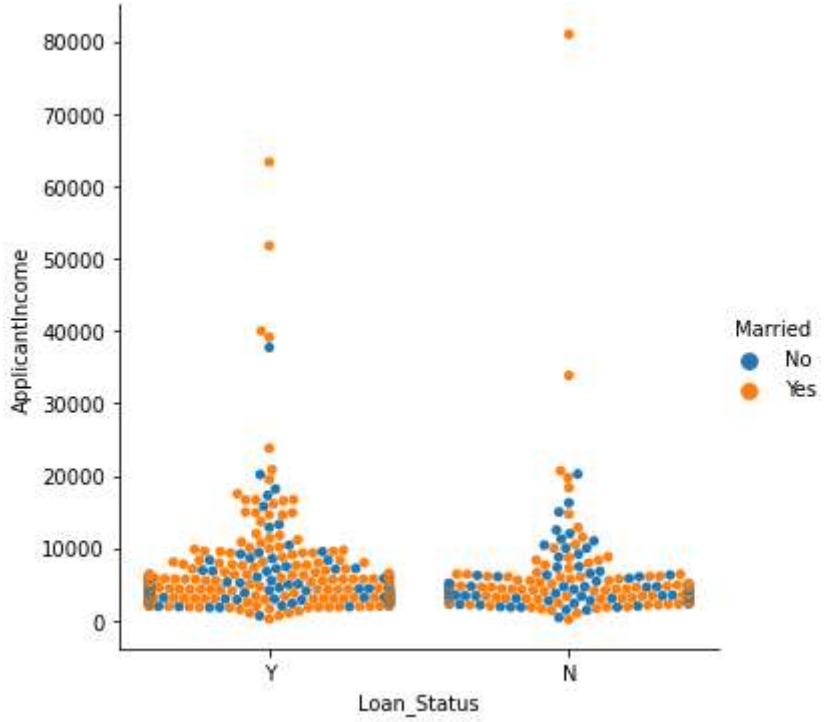

Fig. 4. Loan Approval Status for Marital status with Applicant income

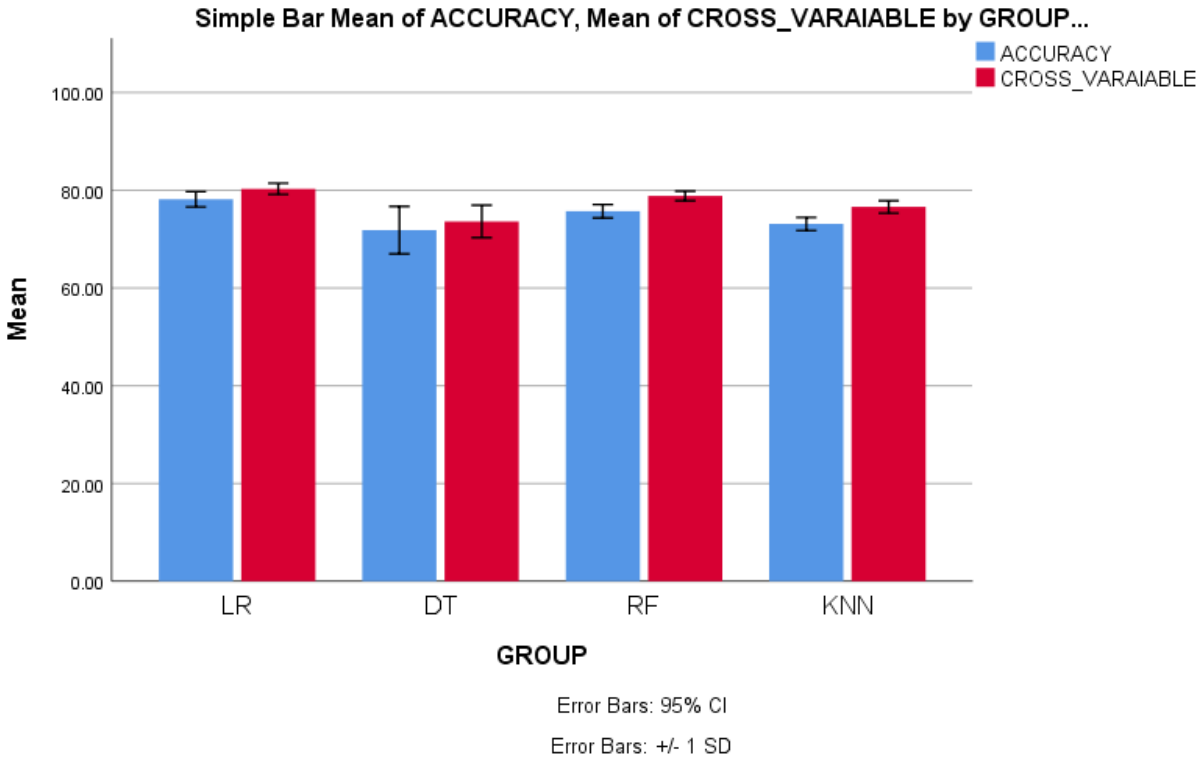

Fig. 5. Comparison of mean accuracy and mean cross validation of LR, DT, RF and KNN algorithms. LR appears to produce more consistent results with higher cross validation and accuracy. X-axis: LR vs DT vs RF vs KNN. Y-axis: Mean Accuracy and Cross Validation ± 1 SD.