ICICC-2021

Loanification - Loan Approval Classification using Machine Learning Algorithms

Kanishk Gupta[a,*] , *Binayak Chakrabarti*[b] , *Aseer Ahmad Ansari*[c] , *Siddharth S.Rautray*[d] , *Manjusha Pandey*[e]

School of computer engineering ,Kalinga Institute of Industrial Technology, Deemed to be university, Bhubaneswar, India

kanishkguptakiit@gmail.com * , binayak042000@gmail.com, a3ahmad.kiit@gmail.com, siddharthfcs@kiit.ac.in, manjushafcs@kiit.ac.in

Abstract:
Banks and other financial corporations have been in the business of lending since the past century. An essential requirement to sustain oneself in such a business is meticulous background checking of a customer before approval of loan failing which the institution lending money carries the risk of non payment of dues by the customer. Such dues which are not repaid back are called bad loans which after 90 days of default become Non Performing Assets (NPA) . According to the Reserve Bank of India ,total NPAs in India are set to reach a twenty year high this year . Thus , the need for developing better models for deciding whether to approve a loan has never been greater . The research work in this paper aims to address this issue and help lending institutions avoid bad loans. A complete background check of each customer has been ensured by feeding entire customer details to our Machine Learning model unlike previous works which relied mostly on specific items like the salary of the client . Ensemble learning models have been the primary focus of this work as Random Forest , Gradient Boosting , XGBoost etc have been used . Out of all the models , the best results were from Gradient Boosting classifier. Finally a Voting classifier has been built by ensembling the aforementioned algorithms . That helped improve the metrics like F1-score and roc- auc-score marginally.

Index Terms- Machine Learning , Ensemble Learning algorithms, classification algorithms

1. Introduction

The current crisis scenario has led to many financial institutions worldwide taking necessary measures to avoid the risk of not being able to recover the money lent to customers . The rampant default of consumer debt has led to many experts re-think whether the current standards and practices are viable enough to secure the company from such events. Thus , it is the need of the hour to use the state of the art machine learning techniques in this domain to predict whether a certain loan to a certain customer should be approved or not.. Most of the traditional practices relied on customer income and/or age as deciding factors for approving loans . However , other aspects like the kind of job the customer is into etc are equally important to be considered . Similarly , it is necessary to check whether the potential debtor already is paying any EMI for any past debts as that will help determine the total EMI he/she will pay after the loan is undertaken . This total EMI can be used to check whether there is any chance of default . In this work various techniques have been used to analyse the importance of each feature of the customer in the dataset . Also, outliers have been detected and scaled appropriately so that the model can be reliable even if future data consists of patterns different from the training data .

Using ensemble learning algorithms gives us the advantage of minimizing errors as those models combine several individual models into one by different methods. The best performing model was Gradient Boosting classifier with Random Forest classifier being a close second . In order to get the best of the top two performance models , a Voting classifier is used which uses the model with the maximum votes with the individual models being Gradient Boosting and Random Forest . The F1 Score of this final model is 0.76 and roc-auc-score is 0.74 . These metrics give a representation of whether the model is performing well in predicting both "approved" and "non approved" categories . With good metrics of our final model , we hope our work will prove to be fruitful and reliable in solving the problem of loan approval in the financial industry .

2. Basic Concepts

The entire work has been done using python 3.8.3 which is a free open source software . The loading and preprocessing of the data has been implemented using pandas and numpy libraries . Pandas is used to check the statistical metrics like mean , deviation etc of the data and also to view it in the form of pivot tables . Numpy is used to transform the data using its arithmetic operations . The loaded data was visualised using matplotlib and seaborn , both of which are libraries having built in functions for various kinds of plots .

The Scikit-Learn library has been extensively used for feature scaling , feature engineering , model development and analysis . The library contains several packages , each serving a particular function of the four mentioned above . The preprocessing and scaling packages help in Feature scaling and engineering . The metrics package helps in analysing the performance of the machine learning algorithms using classification reports , f1 score , accuracy and roc-auc-score [8-15] .

3. Literature Review

The various research works earlier published in this arena have mostly focussed on comparing various machine learning algorithms like Logistic Regression , Decision Tree etc to extract the best one . Few works can be seen which utilize the powerful ensemble learning algorithms which combine several learners into one and have less chance of overfitting compared to the others .

Table 1 compares the works by various authors in this regard.

Research paper by :	Technology used	pros	cons
Tejaswini et. al [1] , Kumar et. al [2], Bhagat et. al [7]	Machine Learning Classification algorithms	Various algorithms were tested and compared and the Decision Tree gave the best results out of all .	Their model was not able to classify one of the labels appropriately . Precision and recall for that label is low as seen from their classification report.
Vaidya et. al [3] , Sheikh et. al [4] , Rath et. al [5]	Logistic Regression	The predictive model , Logistic Regression was fully used and well utilized to classify loan approval .	Other more powerful machine learning algorithms were not used.
Karthiban et. al [6]	Machine Learning Classification algorithms	The performance of algorithms were analysed using metrics like f1 score etc.	However , the metrics were meant for individual classes not the whole data .

Table 1 : Comparison table of various research works

4. Proposed Model / Tool

The research work outlined in this paper aimed to build a reliable model to predict loan approval . This model was constructed with the help of ensemble learning algorithms . Figure 1 shows the research architecture of the proposed model . The voting classifier in this figure is our final model.

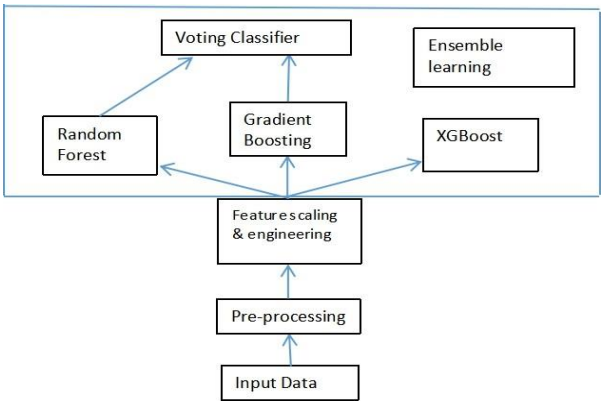

Figure 1 : Research Architecture

5. Implementation and Results :

The initial analysis of our dataset showed a good number of null values in the data . A thorough analysis of the records having null entries proved that the missing values were not at random . Thus , it was necessary to fill them up with appropriate values as removing them from the data might lead to erroneous predictions . The missing entries were filled up using KNN or K-Nearest Neighbours . KNN is an algorithm which estimates the value of a variable using its nearest neighbors . The number of nearest neighbors to be taken into consideration is denoted by hyperparameter K .

Next , it was detected that the dataset was imbalanced . This would have caused a serious threat to our work as imbalanced datasets , when fed into models , might give good accuracy but will fail to accurately predict the minority class samples . Since this was a binary classification problem, class imbalance was easily removed by taking a reduced dataset wherein nearly equal samples of majority class and minority class were present . The subset of the majority class samples was chosen from the original dataset using KMeans clustering algorithm which divides given data into various clusters . This ensured all the clusters of the majority class remain in our reduced dataset thus making it a true representative of the original data.

The exploratory data analysis (EDA) of the new dataset showed the skewness of some of the features like “loan amount “ , “interest rate” etc . Also , boxplot analysis gave us the amount of outliers present which could potentially hurt the model . This issue was treated using feature scaling techniques in which every technique was analysed using probability distribution , QQ plot and box plot . The aim was to find a suitable transformation of the feature which will make the probability distribution close to normal , points of the QQ plot closer to the straight line and have least outliers in the box plot .

Next , the following ensemble models were used : Random Forest - creates several instances of the training set and trains a decision tree on each instance . Finally , the predictions of each tree are ensembled while working on the test set . Gradient Boosting , further , ensures that for each new tree the incorrect predictions of the previous trees are given higher weights. This helps reduce incorrect predictions of the model . XGBoost or Extreme Gradient Boosting has further enhancements like pruning of individual trees which regularises the model thus preventing overfitting .

The receiver operating characteristics curve (roc-curve) for Random Forest , Gradient Boosting and our final model was plotted . This helped visualize the performance of the best performing models in this data . Figure 2 represents the roc curve for Gradient Boosting, Figure 3 represents for Random Forest and Figure 4 represents for our final model or the Voting Classifier respectively.

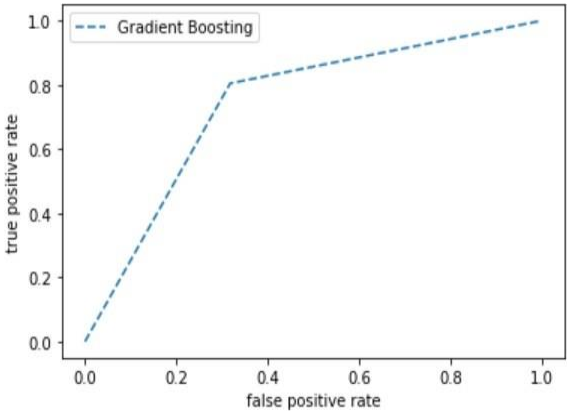

Figure 2 : ROC Curve : Gradient Boosting

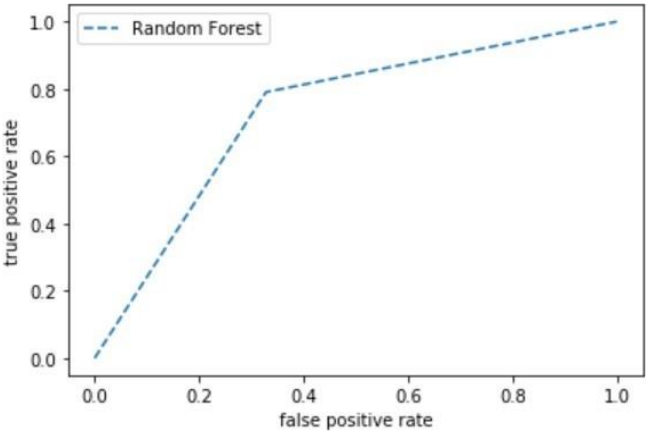

Figure 3: ROC Curve : Random Forest

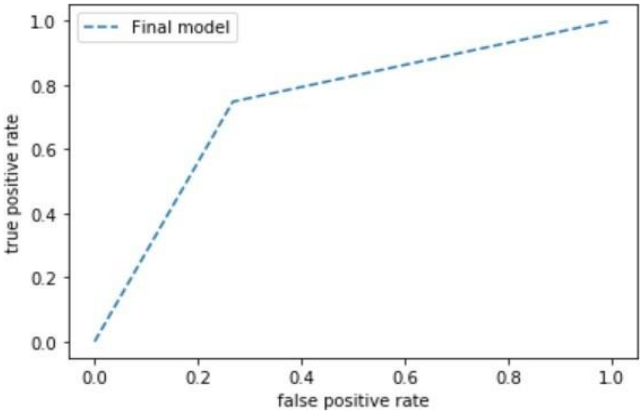

Figure 4 : ROC Curve: Final Model (Voting Classifier)

6. Conclusion

To conclude , our work is aimed to benefit the lenders and the financial industry at large to overcome the potential ramifications of consumer debt becoming a non performing asset . The advancements in predictive algorithms like machine learning have helped solve some of societies’ biggest challenges and is poised to work out this one too .

Acknowledgements

This paper would not have been possible without the support of Dr.Manjusha Pandey and Dr. Siddharth S.Rautray. They inspired us to take this project.

I wish to extend special thanks to mr. Aseer ahmad Ansari who provided us good recommendations with his deep insights and understandings of the topic.

REFERENCES

[1] Tejaswini, J., T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni, and Venkata Rao Maddumala. "Accurate Loan Approval Prediction Based On Machine Learning Approach." Journal of Engineering Science 11, no. 4 (2020): 523-532.

[2] Arun, Kumar, Garg Ishan, and Kaur Sanmeet. "Loan Approval Prediction based on Machine Learning Approach." IOSR J. Comput. Eng 18, no. 3 (2016): 18-21.

[3] Vaidya, Ashlesha. "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval." In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2017.

[4] Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. "An Approach for Prediction of Loan Approval using Machine Learning Algorithm." In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 490-494. IEEE, 2020.

[5] Rath, Golak Bihari, Debasish Das, and BiswaRanjan Acharya. "Modern Approach for Loan Sanctioning in Banks Using Machine Learning." In Advances in Machine Learning and Computational Intelligence, pp. 179-188. Springer, Singapore, 2020.

[6] Karthiban, R., M. Ambika, and K. E. Kannammal. "A Review on Machine Learning Classification Technique for Bank Loan Approval." In 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6. IEEE, 2019.

[7] Bhagat, Abhishek. "Predicting Loan Defaults using Machine Learning Techniques." PhD diss., California State University, Northridge,2018.

[8] Le, DN., Parvathy, V.S., Gupta, D. et al. IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification. Int. J. Mach. Learn. & Cyber. (2021). https://doi.org/10.1007/s13042-020-01248-7

[9] Anupama, C.S.S., Sivaram, M., Lydia, E.L. et al. Synergic deep learning model–based automated detection and classification of brain intracranial hemorrhage images in wearable networks. Pers Ubiquit Comput (2020). https://doi.org/10.1007/s00779-020-01492-2

[10] Shankar, K., Sait, A. R. W., Gupta, D., Lakshmanaprabu, S. K., Khanna, A., & Pandey, H. M. (2020). Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model. Pattern Recognition Letters, 133, 210-216.

[11] Mohanty, S. N., Ramya, K. C., Rani, S. S., Gupta, D., Shankar, K., Lakshmanaprabu, S. K., & Khanna, A. (2020). An efficient Lightweight integrated Blockchain (ELIB) model for IoT security and privacy. Future Generation Computer Systems, 102, 1027-1037.

[12] Pustokhina, I. V., Pustokhin, D. A., Kumar Pareek, P., Gupta, D., Khanna, A., & Shankar, K. Energy-efficient cluster-based unmanned aerial vehicle networks with deep learning-based scene classification model. International Journal of Communication Systems, e4786.

[13] Arora M. and Kansal V., "Character Level Embedding with Convolution Neural Network for Text Normalization of Unstructured Data for Twitter Sentiment", Social Network Analysis and Mining, 9(1), Springer, DOI 10.1007/s13278-019-0557-y, 2019

[14] Pandey S. and Kansal V., "Social Media Analytics: An Application of Data Mining" in Book titled "Data Mining in Dynamic Social Networks and Fuzzy Systems", IGI Global, DOI: 10.4018/978-1-4666-4213-3.ch010 , pp. 212-228, 2013

[15] Arora, M & Kansal, V., "A Framework of Informal Language: Opinion Mining", International Conference on Computing, Communication and Automation", IEEE, pp 41-45, DOI 10.1109/CCAA.2015.7148368, 2015