Solving the problem of loan default problem in the banking sector using machine learning

Deepank Shokeen , Vibhor Grover and Vansh Verma

Netaji Subhas University of Technology

ABSTRACT

In the banking industry, loan default is a still a major issue faced by the banks. It occurs when a borrowers is not able to repay the loan amount, resulting in loss of capital and reputation of the banks as banks use their customers money to give loans to other people. To prevent loan default, lenders typically perform thorough credit checks on potential borrowers to assess their ability to repay the loan and is still an evolving field and an area of research .In this paper, we try to utilize data mining pre processing methods combined with classification based machine learning approaches like KNN, XGBoost, Random Forest, Logistic Regression and Decision Trees to present a high performing solution to this problem. This would help banks reduce the number of steps and resources needed to determine loan eligibility, allowing them to conserve their workforce and financial resources. It turned out that XGBoost and different models show very little distinction on basis of the domain understanding and analysing different performance metric we present our solution comparing with other existing classification models.

1.Introduction

Loans play a crucial role in our lives and can drive the growth of consumption and the economy[1]. Due to imbalance nature of finances in our society, organizations and individuals take loans/credit from banks or other sources to meet their needs. And it has been beneficial for both the lender and the borrower. However, it is important to accurately assess whether a borrower is eligible for a loan. As the number of bad debts increase the finances of the banks start to dwindle which could lead them and the complete economy into a crisis. We have used data processing along with classification machine learning algorithm to predict potential defaulters, helping banks make better decisions in the future. By analysing this data, banks can more accurately identify borrowers who are at risk of default, allowing them to make more informed lending decisions and reduce the risk of financial crisis. Overall, the use of these algorithms can help banks better manage their loan portfolios and support the growth of the economy.

In the past, manual review—which required a lot of time and manpower used to be to analyse loan default. Though nowadays, banks have started using new age technology involving machine learning, data science approaches to automate the prediction of loan default, which can improve the accuracy and efficiency of the process. The abundance of online shopping and mobile payments has provided banks with a vast amount of transactional data, and machine learning models have demonstrated successful applications in a variety of fields[2]. These factors have encouraged the banking industry to adopt machine learning for predicting loan default.

Using various pre-processing data mining steps, generating new metrics experimentally and combining with a classification algorithm XGBoost[3] we were able to come up with a well performing outcome. We compared our framework with 4 of the foremost used classification models currently used in the industry which include: KNN [4], Decision Tree [5], Logistic Regression [6], Random

Forest[7]. Using the public information available for default prediction, by carefully examining the potential risks and performance of the model, we can better understand its limitations and potential for success in predicting loan default.

2. Related work

In this section, we will review some recent researches on loan default prediction.

For default prediction, by minimizing the process required to determine whether a loan is eligible, banks can save human and financial resources. Authors of [8] propose more efficient strategies using machine learning based classification methods such as Logistic Regression, SVM, Decision Tree, Random forest, and Artificial Neural Networks. Metrics such as log loss, Jaccard similarity factor, and F1 score to measure the accuracy of these approaches were proposed. Comparison of these measures to evaluate the accuracy of the predictions was done.

Improved predictive accuracy is linked to the application of new machine learning classification algorithms for loan default prediction [9], but it also brings new model risks, particularly with regard to the supervisory review procedure. Recent industry studies frequently point out that the lack of clarity regarding how regulators can evaluate these risks can serve as a barrier to innovation. To assess the effectiveness of various machine learning techniques, the authors of [9] provide a unique methodology for quantifying model risk modifications in the study.

The authors of [10] aim to classify whether people are able to pay off their debt while preventing banks from incurring significant losses. If a borrower defaults on a large loan, it could bankrupt a bank and potentially lead to a financial crisis. The number of defaults would naturally increase as the number of borrowers increases. The study compares and contrasts three machine learning techniques: Random Forest, Logistic Regression, and XG Boost, in order to determine which model provides the most accurate performance.

Two key queries are frequently posed in the banking sector: (1) Is there high chance of default by the borrower? (2) Shall the loan be given to the borrower given the risk associated?[11]. In order to help banking authorities by analysing certain features to predict whether a loan should be issued and by making it easier to choose acceptable loan applicants from a given list, authors of [11] offer two machine learning models in response to these questions. The article compares and contrasts decision trees and random forests as two algorithms. The same data set is used to test both methods, and it is determined that the random forest approach performs better and is more accurate than the decision tree algorithm.

Another study focused on modelling and predicting the willingness to repay a credit card loan[12]. Methods used in that study include machine learning. 11 variables were analysed and the performance of five methods was compared with ROC and AUC scores. The results of the study were. The random forest method was considered the best for processing the basic credit card dataset with an AUC of 89%. That model can contribute to the resolution of possible defaults and is a huge boon to the credit card industry. From the manager's point of view based on the PDP, it can be judged that the higher the income and credit card limit of 7 million to 50 million won, the higher the possibility of default.

Another study compared the effectiveness of machine learning models for forecasting underlying risk to that of conventional statistical models [13]. The study discovered that when there is little information available, like in the case of external credit risk assessment, machine learning models had superior discriminant power and accuracy than statistical models. However, this advantage is

minor when the data collection is limited and when sensitive information, such as credit behaviour indicators, is also available. The study also applied machine learning score-based credit distribution algorithms. Overall, the study discovered that by concentrating loans on riskier, bigger borrowers, machine learning models can assist lower lenders' credit losses.

3.Methodology

3.1. Dataset Preparation

The dataset used to build the model is a real world data publicly available on Kaggle. It includes loan data with 34 columns (features) and 174000 rows. The data was in the form of a CSV file.

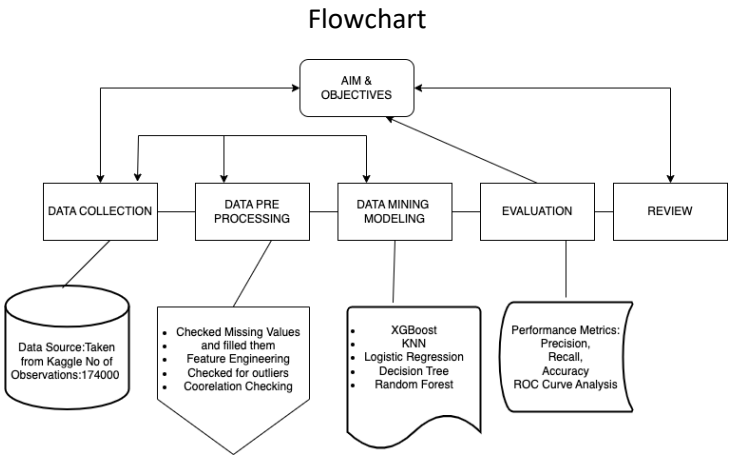

Figure 1. Overview of all the stages

In Figure 1. the flowchart explains the complete process of our methodology from data collection to data pre-processing to data mining modelling then evaluation and reviewing our complete process. The details of the following processes is as follows:

3.2. Data Pre processing

Data Pre Processing is the transformation of the raw data into more meaningful and efficient format which would help the machine learning models to perform better [14]. The real world data is generally full of noisy data,with missing values and duplicate data but for machine learning algorithms to perform efficiently we need the data to be in processed form which is the aim of this step.

We performed the following pre-processing steps on our data.

i. We removed the unimportant features which did not provided any viable information.These features were id of the borrower, year of borrow as all the data was from the same year, loan limit,gender as there was much of correlation found, approval in advance , credit worthiness, open credit, interest_rate_spread, neg ammortization,construction type,occupancy type,secured by, submission of application, ltv ,and a few more unnecessary features.
ii. We had to fill the missing values to provide a consistent data so that our model is able to perform accurately so we filled the missing values according to the features. Most of the features missing

values were filled by taking the mean of the non-null values and the missing values of the feature 'age' was filled with the mode(most frequent) value.

iii. We had some outlier values for 'income' feature which was found using data visualization these values were then replaced by the mean income value.

iv. We also had outlier with 'interest rate' feature where it was zero which was also then replaced by the mean interest rate values.

v. We had many categorical features in our data, these were converted into numerical data by encoding. These features include 'loan type' and 'age'.

vi. We also introduced a new metric to better utilize the multicollinearity of the features.

$$LRP\ Score = \frac{Loan\ Amount * Property\ Amount}{Rate\ of\ Interest}$$

$$LC\ Score = \frac{Loan\ Amount}{Credit\ Score}$$

vii. Synthetic minority oversampling technique (SMOTE) was done to solve the imbalance nature of our dataset.

viii. Data partitioning was then done to train our ML model.

Features	Description
Loan_type	Type of loan applied for
Loan_amount	Value of the Loan
Rate_of_Interest	Interest Rate applicable for the loan
Term	Duration of the Loan
Property_value	Collateral property value
Income	Income of the applicant
Credit_score	Credit score of the applicant
Age	Age of the applicant
Status	Current loan status
LRP Score	Combination of Loan amount, property amount and rate of interest
LC Score	Combination of Loan amount and Credit Score
Dtir1	Debt to income ratio

Table 1.Features after Pre-Processing

Table 1. shows all the features along with their description which are a result of the data pre-processing stage. These attributes would be used to predict the final outcome.

3.3. Model Training

In our experiment, we would we using XGBoost machine learning algorithm to train our model. The dataset was partitioned into two parts one containing 80% of the entire dataset and other containing 20%. The 80% part is the training part and rest is for testing purpose. We trained the model using the training dataset and evaluated its performance using the testing dataset.

XGBoost stands for Xtreme Gradient Boosting, it is an algorithm based on decision trees classification model that improves upon the gradient boost algorithm. This technique generalizes the component

classification decision trees by using an arbitrary differentiable loss function for model optimization.Steps involved in XGBoost algorithm are:

i. First, we input the training set {(xi,)}, learning rate *B*, loss function $Lf(y,P(x))$and. weak learners *M*.

$$\hat{p}_{(0)}(x) = \arg min_{\theta} \sum_{i=1}^{N} Lf(y_i, \theta) \quad (1)$$

ii. Use a constant to initialize the model

For *k* = 1 to *K*:

iii. Calculate hessians(t) and gradient (q):

$$\hat{q}_k(x_i) = \left[\frac{\partial Lf(y_i, p(x_i)}{\partial p(x_i)}\right]_{p(x) = \hat{p}_{k-1}(x)} \quad (2)$$

$$\hat{t}_k(x_i) = \left[\frac{\partial Lf(y_i, p(x_i)}{\partial p(x_i)^2}\right]_{p(x) = \hat{p}_{k-1}(x)} \quad (3)$$

iv. Train a base learner on the training set.$\left\{x_i, \frac{-\hat{q}_k(x_i)}{\hat{t}_k(x_i)}\right\}$ by solving the optimization question below:

$$\hat{\varphi}_s = \arg min_{\varphi} \sum_{i=1}^{N} {}^{1}/_{2}\, \hat{t}_k(x_i) \left[-\frac{\hat{q}_k(x_i)}{\hat{t}_k(x_i)} - \varphi(x_i)\right]^2 \quad (4)$$

$$\hat{p}_k(x) = B\, \hat{\varphi}_k(x) \quad (5)$$

v. Update the model:

$$\hat{p}_k(x) = \hat{p}_{(k-1)}(x) + \hat{p}_k(x) \quad (6)$$

vi. Output:

$$\hat{p}(x) = \hat{p}_{(x)}(x) \sum_{k=0}^{K} \hat{p}_k(x) \quad (7)$$

3.4.Results And Evaluation

The dataset is enormous & consists of multiple deterministic factors like borrower's income, age, credit score. The dataset is subject to strong multicollinearity & empty values that is why we tried feature selection and event making some new features by combining the existing ones.

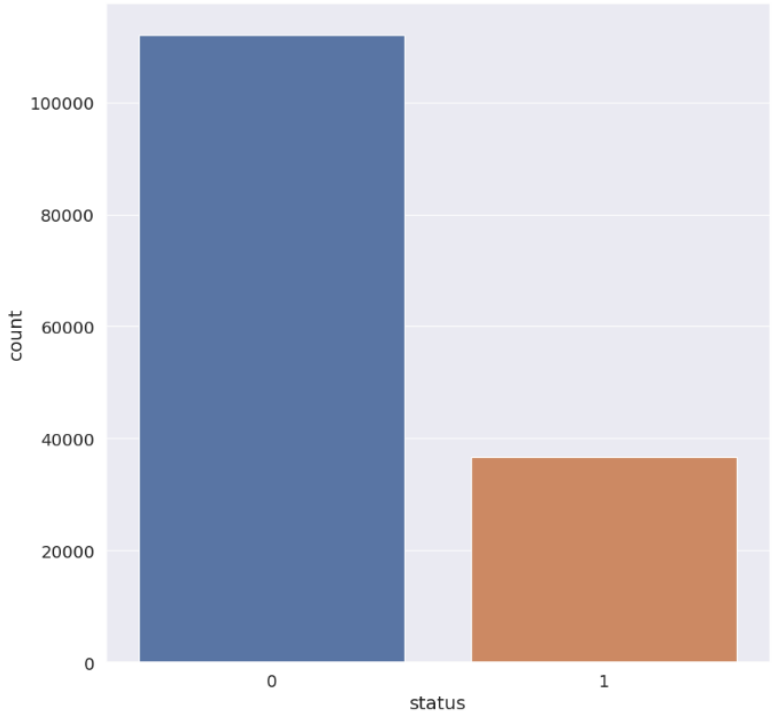

Figure 2. Bar graph of distribution of status

From figure 2. We can observe that our dataset is imbalanced as we have more number of people not defaulting and less number of people defaulting.

We applied various common classification algorithms and our framework of data preprocessing combined with XGBoost algorithm to have a comparative study. We calculated various performance metrics [15], accuracy and ROC Curve to have a detailed looks at the outcomes.

Performance Metrics:

Precision: It is the measure of how many of the positive values are predicted correct.

$$Precison = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad (8)$$

True Positive means the sample which are likely to default are classified as likely to default as well.

False Positive means the sample which are not likely to default are classified as likely to default.

Recall: It is the percentage of the true positives that were predicted correctly.

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \quad (9)$$

False Negative means the sample which are likely to default are classified as not likely to default.

Accuracy: It is the percentage of correctly classified data instances over the total number of instances.

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \quad (10)$$

True negative means the sample which are not likely to default are classified as not likely to default as well.

	Accuracy	Precision	Recall
XGBoost	**97.77**	**93.7**	**97.59**
Decision Tree	96.87	93.3	94.24
Random Forest	96.2	91.5	93.4
KNN	80.5	62.99	54
Logistic Regression	49.16	25	51.74

Table2. Performance Metrics of several classification Models

We can observe from Table2 that our framework with XGBoost is outperforming all the others in terms of the accuracy. But our problem of banking loan default generally contains an imbalanced dataset with less no of people defaulting and more no of people not defaulting so in such scenarios we prefer Precision and Recall as more important performance metrics. Our preference would be to have false negatives as less possible as banks would not desire to have a person likely to default classified as not defaulting and Recall is the performance metric which penalizes when a false negative occurs so Recall is our main evaluation metric and our framework with XGBoost had significantly better Recall than others.

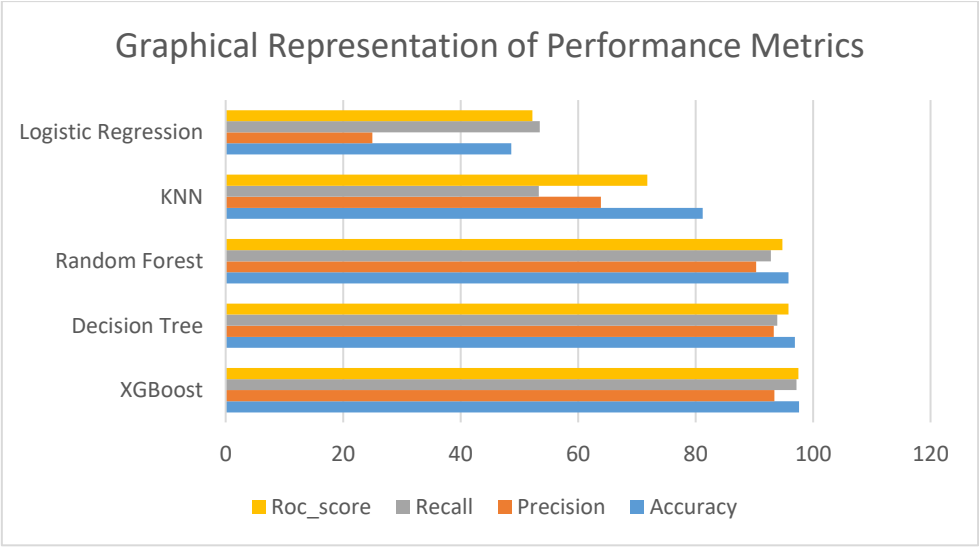

Figure3.Bar Chart of Perform Metrics

In Figure 3 Blue bars represent the Accuracy of all the models, Orange represents the Precision, Grey represents the Recall and yellow bars represents the ROC scores.

ROC Curve: Receiver Operating Characteristics is a plot of the probability of a true positive against the probability of a false positive for all possible threshold values.

Let's also have a look at a ROC curve including all our ml models as well.

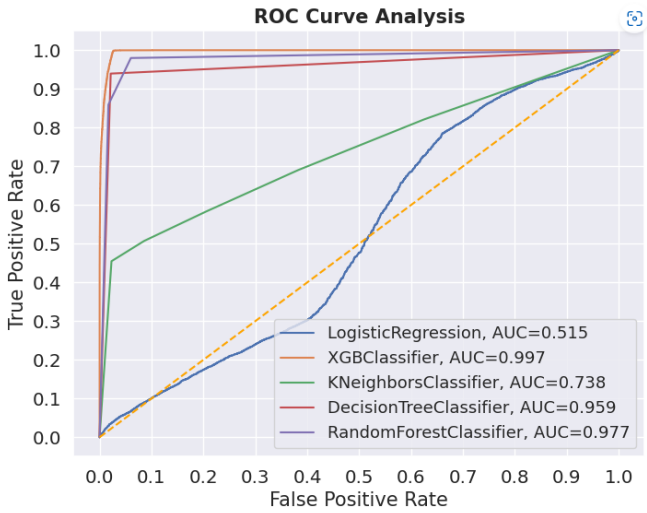

Figure4.ROC curve of all ML models

In Figure 4, the x-axis represents the false positive rate and the y-axis represents True Positive Rate, and the area under the curve represents the measure of usefulness of a test. The orange curve represents XGBoost Model, purple curve represents Random Forest Model, red curve represents Decision tree Model, green curve represents KNN Model and blue curve represents Logistic Regression Model.Logistic Regression model did not perform well for our dataset as clearly we can see for none of the threshold is the true positive rate high and false positive rate low leading to an S shape type of an arbitrary curve .Xgboost, Random forest and Decision tree performed really well.

From figure 4, We can see that the XGBoost model's ROC Curve is closest to the upper left corner and has the largest area under the curve, indicating the best performance as we were able to get high positive rate and less false positive rate.

4. Conclusion

Loan default problem is still one of the most challenging and important problem in the banking industry for their fluent working. In our report, we proposed a solution to the loan default problem by deducing new parameters and combining it with XGBoost algorithm.

We compared our framework to various other classification learning algorithms which included K Nearest Neighbour classification, Decision Tree Classification, Logistic Regression and Random Forest classification and on the basis of domain based performance analysis we found that our algorithm outperformed all of them. Therefore, concluding that our model was able to provide a solution which would help banks overcome the issue of loan default hence helping them maintain a healthy financial system.

References

1. Lai, Lili. "Loan default prediction with machine learning techniques." In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pp. 5-9. IEEE, 2020.

2. Wang, Yuelin, Yihan Zhang, Yan Lu, and Xinran Yu. "A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data." *Procedia Computer Science* 174 (2020): 141-149.

3. C Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016

4. Guo, Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. "KNN model-based approach in classification." In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 986-996. Springer, Berlin, Heidelberg, 2003.

5. Hauska, Hans, and Philip H. Swain. "The decision tree classifier: design and potential." In *LARS Symposia*, p. 45. 1975.

6. Sperandei, Sandro. "Understanding logistic regression analysis." *Biochemia medica* 24, no. 1 (2014): 12-18.

7. Ho, Tin Kam. "The random subspace method for constructing decision forests." *IEEE transactions on pattern analysis and machine intelligence* 20, no. 8 (1998): 832-844.

8. Aditya Sai Srinivas, T., Somula Ramasubbareddy, and K. Govinda. "Loan Default Prediction Using Machine Learning Techniques." In *Innovations in Computer Science and Engineering*, pp. 529-535. Springer, Singapore, 2022.

9. Alonso Robisco, Andres, and Jose Manuel Carbo Martinez. "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction." *Financial Innovation* 8, no. 1 (2022): 1-35.

10. Ramesha, Nishanth. "Machine Learning Based Approaches to Detect Loan Defaulters." In *International Conference on Advances in Computing and Data Sciences*, pp. 336-347. Springer, Cham, 2022.

11. Madaan, Mehul, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. "Loan default prediction using decision trees and random forest: A comparative study." In *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012042. IOP Publishing, 2021.

12. Sayjadah, Yashna, Ibrahim Abaker Targio Hashem, Faiz Alotaibi, and Khairl Azhar Kasmiran. "Credit card default prediction using machine learning techniques." In 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), pp. 1-4. IEEE, 2018.

13. Moscatelli, Mirko, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. "Corporate default forecasting with machine learning." *Expert Systems with Applications* 161 (2020): 113567..

14. García, Salvador, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Vol. 72. Cham, Switzerland: Springer International Publishing, 2015.

15. Canbek, Gürol, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. "Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights." In *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 821-826. IEEE, 2017.