AN APPROACH TO LOAN APPROVAL PREDICTION USING MACHINE LEARNING

B. Yamuna ,B.Tech Student, Department of Computer science and engineering, Vignan's Institute of Information Technology, Visakhapatnam AP., INDIA.

Ch. Praneeth ,B.Tech Student, Department of Computer science and engineering, Vignan's Institute of Information Technology, Visakhapatnam AP., INDIA.

D. Sai Nithin ,B.Tech Student, Department of Computer science and engineering, Vignan's Institute of Information Technology, Visakhapatnam AP., INDIA.

E. Sri Ramya ,B.Tech Student, Department of Computer science and engineering, Vignan's Institute of Information Technology, Visakhapatnam AP., INDIA.

Chintam Anusha , Assistant Professor , Department of Computer science and engineering, Vignan's Institute of Information Technology, Visakhapatnam AP., INDIA.

Abstract

Now a days banks are receiving lakhs of loan applications daily. This is delay in approval of loan to the customer as humans need to verify each ofthem manually. So even after havingenough resources to approve loans it is bigproblem for banks to approve applications in time. In order to resolve this problem, weare planning to build an ML application which can reduce the time required to approve a loan using ML based prediction model to approve the loan with minimal human intervention by filtering huge number of applications and forward very few applications for human verification.For this we can use several popularmachine learning algorithms and predictionmodels. For this we are testing with Decision tree, Logistic Regression, Random Forest tree, Support Vector machine and XGB.

Keywords: Prediction, Machine Learning, Logistic Regression, Decision tree, Support Vector machine, Random Forest tree and XGB.

Introduction

Banks have various products to sell in our banking system, but their major source of money is their credit lines. As a result, theycan profit from the interest on the loans theycredit. Loans play a key role in determiningprofit or loss of the bank, i.e., whether consumers repay the loan or fail. Any bank can avoid its Non-Profiting Assets by pre-identifying loan absconders. As the outcome,research into this process is crucial. Existing works now have proved that there are so many implementations for studying the topic of loan escape control. However, as perfect forecasts are critical for the profit maximization, it is crucialto research and compare the various methodologies. To explore the topic of predicting loan escapers, the XGBoost model is utilized, which is a much important procedure in predictive research. Kaggle data is used to research and predict. The several performance metrics were computed using XGBoost models. Performance indicators like sensitivity and specifications were used for comparing these models. The final results show that the model gives-out different outcomes. Model is slightly better as it also includes attributes(customer personal attributes suchas age, dependents, education, background, employment, and etc) other than just financial information (which indicates a customer's money background only) that should be taken when calculating the probability of loan escape correctly. As a result, by calculating the possibility of loan escape, the ideal customers to target for loan giving will be easily identified using a XGBoost model approach. The model concludes for a bank must not only focus on wealthy consumers when giving loans to applicants, but we should also take consideration of a customer's remaining attributes, which play an essential factorin credit repayment and forecasting loan escapers.

Existing work

In existing system, the whole process of determining the outcome of loan application is Purely based on the financial aspects of the applicant such as the income, Net worth, credit background etc. this might not be ideal for every scenario as many factors which show a major impact on credibility are totally neglected. There could be many situations where the family background and educational background ofthe application can over-shadow the financial lacking of the applicant.

Methodology

This Our model provides solution to the loan approval in much better way when compared to existing models. Instead ofconsidering only the financial factors of theapplicant we also consider the non- financial factors which can impact the credibility of the applicant which makes our solution much ideal for every scenario. Forthis we are using additional factors such as gender, marital status, number of dependents, educational qualification,employment status etc.

Data collection

For our solution the data has been collectedfrom Kaggle one of the widely used datasetproviders for machine learning which is further split in training ang testing sets. Out of which the training set is used to train themodel and testing set is used to calculate theaccuracy of the trained model.
Our dataset consists of 1010 records and 13 attributes shown below

- Loan id
- Gender
- Married/Not Married
- Family Size
- Qualification
- Employment type
- Applicant Income
- Co-applicant background
- Required amount
- Loan span
- credit score
- property area
- result

Model training

To find the best fitting model, we aretraining five different machine learning model with our data and analyzing their results based on different metrics.

Fig. 1 Flow Diagram of proposed work

Logistic Regression

Logistic regression predicts the result class of a single class depending on variable. So, the output must be a non-numerical or numeric value. It could be anything yes or no, 0 or 1, True or False, etc. but rather than giving the directly as 0 and 1, it gives the decimal values that range between 0% and 100%.

Random forest tree

Random Forest tree is the model that consists of a group of prediction trees on different sub-divisions of an input data and finds its mean to improvise the accuracy of the data given. Instead of depending on only one decision tree model , this random forest tree considers the prediction as multiple trees and depending over the strength of classifications, and that gives the ultimate outcome.

Extreme Gradient Boost

XG-Boost is an improvised gradient boosting method designed for being more efficient**,** modular, and reusable **.** It implements machine learning codes using the simple gradient boost framework. XG-Boost provides a multiple tree-like boosting technique that solves multiple data-based problems in a quick and precise way. The same unique code works on many distributed environments in real-time and can solve problems beyond thousands and thousands of samples.

Decision-tree classifier

Decision-Tree is a Supervised method which is usually used for any of non-numeric and numeric problems, but highly it is used for solving non-numeric based problems. that is just tree like classifier, in which non-leaf nodes represent the features of a given data, edges represent the decision conditions and every ending node represents the output class. While implementing this algorithm, the main problem is that how to choose the best attribute for the root node and for sub-nodes. So, to solve this kind of problem there is a technique which is called as Attribute selection measure.

Support Vector Machine

Support Vector Machine is a supervised learning model mostly used for any classifying and also regression cases. Not so good for regression problems but its best applicable for classification. The aim of this algorithm is for drawing a hyper-plane in multidimensional area that uniquely divides these points.

Results and Discussion

Accuracy score

Accuracy is most basic metric used in machine learning. It is the percentage of size of right predictions made to the size of input samples. Classification accuracy is good, but it gives Fake Positive result of getting more accuracy. The problem occurs due to the possibility of miss-classification of small class samples are too high. From fig.2 xgboost gives best accuracy score

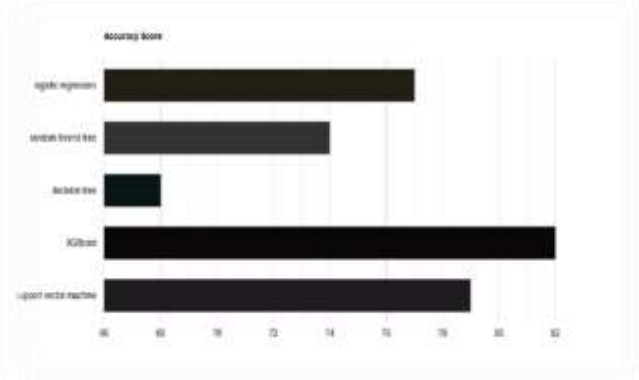

Fig.2.accurecy score

Mean square error

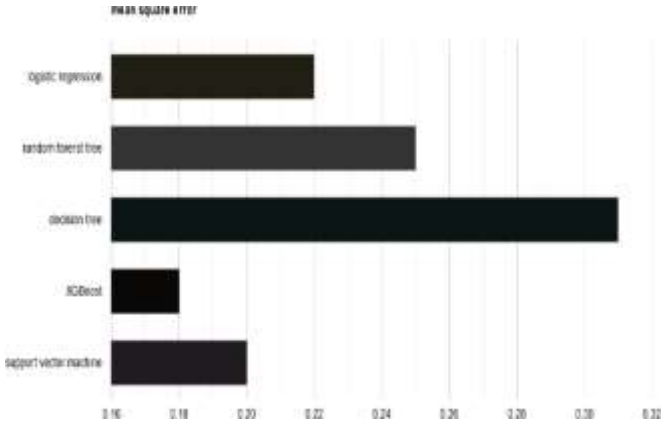

Fig.3.Mean Square Error

It is just like mean absolute error, but the change is it uses the square of avg of difference between predicted and original outputs. The key advantage to take thismeasure is, it is easier to compute the gradient but in the case of mean absolute error it takes complicated calculation tool to find the gradient. By taking the square oferrors it gives larger errors more than smaller errors, so we can look more onlarger errors. From fig.3 XGBoost gives least Mean Square Error

Recall

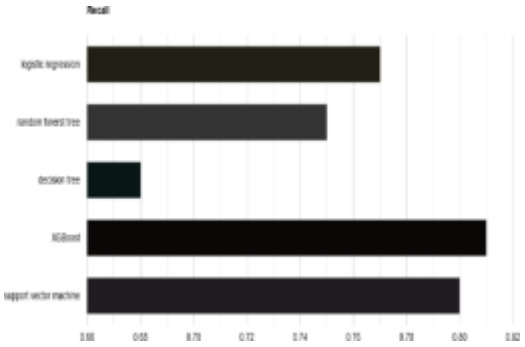

Fig.4.Recall

The recall is considered as ratio of thenumbers of +ve samples rightly predicted as +ve to the whole number of +ve output samples. The recall calculates the model's probability to find the +ve samples. The better the recall, the better +ve samplesdetected from fig.4 XGBoost gives best recall

F1 score

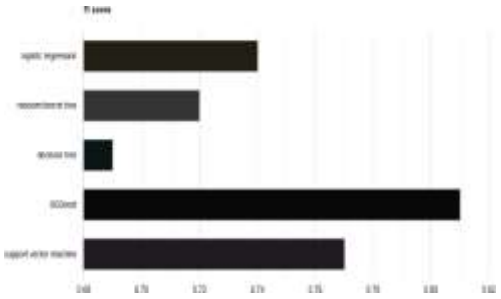

Fig.5.F1 score

The F1 score is described as the harmonic mean of calculated precision to recall,F1 score touches the max score near one and the min near 0. The usage value of the precision and the recall for this F-1 score are same.
The equation to find F1 score is

F1 = 2 x (perc x rec) ÷ (perc +rec)

From fig.5 XGBoost gives best F1 score

From these above graphs that XGBoost is giving the best results in all the

metrices we are using it in our solution . and we can also see that decision tree is giving the worst results.

Conclusion

- As our prediction model uses several attributes of the applicant which also include non-financial attributes we can obtain highly reliable model when compared to the ones which include only financial attributes.

- For our dataset the best accuracy is achieved for the XGBoost model which is 0.82 .

- As we can clearly see in results that XGBoost is best in all metrices we conclude it is best model for our dataset.

References

[1] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue2, Taylors & Francis.
[2] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.
[3] Aakanksha Saha, Tamara Denning, VivekSrikumar, Sneha Kumar Kasera. "Secrets in Source Code: Reducing False Positives using MachineLearning", 2020
[4] InternationalConference on CommunicationSystems &Networks (COMSNETS), 2020.
[5] X.Frencis Jensy, V.P.Sumathi,Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S,November 2018.
[6] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,"Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2,Mar-Apr 2019
[7] Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujrat Research History,Volume 21 Issue 14s, December 2019.

[8] Bing Liu, “Sentiment Analysis and Opinion Mining,” Morgan & Claypool Publishers, May 2012.
[9] Sravani.A, Anusha.C, Shankar.N.V.S “A Comparative Analysis of Machine Learning Algorithms in Stock Prediction”, Proceedings of the International Conference on Industrial Engineering and Operations Management, 2021, pp. 2619–2623.