Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

RF-XGBoost Model for Loan Application Scoring in Non Banking Financial Institutions

Mounika Koduru[1], Pranati Chunduri[2], Manasa Jonnadula[3], M. Phanidhar[4], Dr. Kudipudi Srinivas[5]
[1,2,3,4]IV/IV B. Tech,
[5] Professor
Department of Computer Science
Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada,
A.P, India.

***Abstract*— Non banking financial institutions which in short form known as NBFCs follow a simple four steps process in processing a loan application. These include filling the online application, uploading documents, credit analysis done by the company and disbursement. In this paper, we proposed a hybrid model called Random Forest Extreme Gradient Boosting (RF-XGBoost). In this model, application score is calculated and then status of the application is predicted whether to approve the loan or reject it. Here, a Random forest classifier is used to get the importance of each feature. Further, XGBoost and a simple neural net are used to predict the loan status. This entire lending space in assessing the credit risk is based on Artificial Intelligence technology, making it a profitable situation for borrowers and the platform lenders. Generating the unique score for the loan application will help the lender to process the loan more effectively to the customer. The experimental results showed that, RF-XGBoost model performance is best compared to Logistic Regression, Random Forest, Linear SVM and Neural Net-3 Layers.**

Keywords*— Application score, credit analysis, credit score, Boosting, neural net.

1. INTRODUCTION

Before applying for a loan, it is essential to understand the application process to successfully get your loan to be approved. Loans provided by the organizations are useful for many short terms and long term works in our life. In addition, having a good credit score increases the value of the loan application. Personal credit scoring plays a very crucial role in the risk management of many commercial banks. These credit scoring techniques are capable of reducing the risk management of many banks. With the growth of credit cards and personal loans, credit scoring technology has been widely applied into many lending areas and loan management of banks. Credit score helps the organizations in making the loan approval processing much easier. In the literature Orgler (1970) used linear regression method for assessing the customers loan risk[1].Logistic Regression model was applied by Wiginton (1980) to predict the credit score of the loan application[2].Neural nets solve a lot of problems because of their nonlinear capability[3-4].

This paper first calculates the application score of the customer based on the important features which are selected by using random forest classifiers and then boosting technique and neural nets are applied on the data to predict the loan status of the customer.

2. RELATED WORK

2.1 Application Score

Data analysis is the key step in calculating the application score of the customer. Some of the important features are the customer's annual income, debts, employment length, and loan type. Each bank develops their own scoring model based on the mentioned characteristics by adding some more too.

2.2 Feature importance

Random forest classifier is applied on the sample data for generating the importance of each feature. Figure-1 depicts some of the features which are marked as important from our data. These include employment length, application score, annual income, debts, loan type, interest rate.

The target variable in the data provided is the loan_status which describes whether the loan is approved (1) or rejected (0). Data provided is shuffled randomly. Later the train test split is applied in the ratio 80:20 where the model is built on 80% of the data and tested on the other 20%. Some of the new features are generated from the provided data. These include debt-to-income ratio, loan installment amount, monthly savings and finally application score of the candidate.

```
('loan_status', 0.7213334224355826)
('age', 0.0001816106122744459)
('loan_amount', 0.009289300834287039)
('loan_type', 0.001273436611104788)
('loan_term', 0.000631743901702858)
('interest_rate', 0.00038459879631475935)
('installment', 0.0009139177745434648)
('employment_length', 0.18990562555338666)
('marital_status', 2.465129608991521e-05)
('home_ownership', 0.0002686392051750125)
('total_annual_income', 0.0026046161637916106)
('total_annual_debt', 0.0158594177114285)
('loan_purpose', 0.0019219935216546837)
('debt_to_income_ratio', 0.019972962627352688)
('application_score', 0.03371835259875552)
('monthly_savings', 0.001715710356555443)
```

Fig 1: Important features of loan approval

3. PROPOSED SYSTEM

3.1 Workflow

The algorithm of the proposed model called RF-XGBoost is as follows.

Step 1: Load the sample data using a python environment.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Step 2: Perform Exploratory Data Analysis on the data which include dropping the columns having null values more than 90%, table columns and imputing the missing valued columns.

Step 3: Apply correlation on the data and extract the dependent features among them.

Step 4: Apply random forest classifiers to extract the importance of each feature present.

Step 5: Now split the data into train and test in the ratio of 80:20. Fit the model on the training set.

Step 6: Now apply the XGBoost and Neural Net with 2 hidden layers on the training and test data set and extract the results.

Step 7: Now compare the accuracy of the 2 models built and finalise the better one.

3.2 Extreme Gradient Boosting (XGBoost)

The beauty of the XGBoost lies in its scalability and reliability. It also offers efficient memory usage. It is one of the ensemble models. Ensemble models offer a systematic solution to combine the predictive power of multiple machine learners and provides the final single result of all combined. XGBoost library implements the gradient boosting tree.

3.3 Neural Net with 2 hidden layers

In this model, our input layer consists of 16 features provided as input, two hidden layers and an output layer. Around 100 epochs are used for training the neural net. Optimiser 'adam' is used along with the loss function 'binary_cross entropy' .The activation function in the output layer used is 'sigmoid' whereas in the input and hidden layers, it is 'ReLU (Rectified Linear Unit)'. A batch size of 100 is used in the training.

4. ANALYSIS OF RESULTS

Following parameters are used to analyse the performance of the RV-XGBoost model with Logistic Regression, Random Forest, Linear SVM and Neural Net-3 Layers. A list of around 16 features is used for the model building.

Accuracy: How correctly the classifier is predicting.

(TP+TN)/(TP+FP+TN+FN)

True Positive Rate: How often does the model predict yes when it's actually yes?

TP/actual yes

True Negative Rate: How often does the model predict no when it's actually no?

TN/actual no

The existing systems are built on logistic regression, SVM [5] and Random Forest classifiers. From the figure-2, it is clearly observed that, the newly used RF-XGBoost model has produced better results in the metrics of accuracy and true negative rate.

Figure-2 depicts the performance analysis of the existing models and the proposed RF-XGBoost model. Here it is clearly observed that, the accuracy of the XGBoost is 91 when compared with Random Forest which is 87. So XGBoost is considered as our final model for this loan prediction status.

Name of the Model	Data Set Size	TAP	TAN	TN	TP	FP	TN Rate	TP Rate	Preci-sion	F1-Score	Accu-racy
RF-XGBoost	**110000**	**40138**	**69862**	**66686**	**32415**	**3176**	**7723**	**81**	**91**	**89**	**91**
Logistic Regression	110000	40138	69862	59255	18004	10607	22134	45	68	65	71
Random Forest	110000	40138	69862	66650	28615	3212	11523	71	88	85	87
Linear SVM	110000	40138	69862	53833	12094	16029	28044	31	54	53	61
Neural Net-3 Layers	110000	40138	69862	55165	26016	14697	14122	65	72	72	74

Fig 2. Performance Matrix

5. CONCLUSION

In this paper, we introduced a hybrid model called **RF-XGBoost** to calculate Loan Application Scoring and predict the status of application whether to approve the application or reject. The experimental results showed that, RF-XGBoost model has produced better results and the accuracy of the XGBoost is 91 when compared with accuracy of Logistic Regression, Random Forest, Linear SVM and Neural Net-3 Layers. In future work, extend our proposed system for high dimensional and complex unbiased datasets.

REFERENCES

[1] Orgler, Y. E, A Credit Scoring Model for Commercial Loans, Journal of Money, Credit, and Banking; 1970,2, Iss.4, p. 435-445.

[2] Wiginton, J. C, A note on the comparison of logit and discriminate models of consumer credit behavior. Journal of Financial and Quantitative Analysis; 1980,15,p. 757-770.

[3] Jianping Li, Zhenyu Chen , Liwei Wei, Weixuan Xu and Gang Kou, Feather Selection via Least Squares Support Feature Machine. International Journal of Information Technology & Decision Making; 2007, 6(4),p. 671-686.

[4] Abdou, H, Pointon, J. and El-Masry, A., Neural nets versus conventional techniques in credit scoring in Egyptian banking, Expert Systems with Applications. v35 i3. 1275-1292.

[5] Jianping Li, Gang Li, Dongxia Sun, Cheng-Few Lee, Evolution Strategies Based Adaptive Lq Penalty Support Vector Machine with Gauss Kernel for Credit Risk Analysis, Applied Soft Computing; 2012, 12(8),p. 2675-2682.

[6] Li, J., J. Liu, W. Xu and Y. Shi Support Vector Machines Approach to Credit Assessment, in P. M. A Sloot et al, eds., ICCS 2004, LNCS 2658, Springer, Berlin; 2004, p.892-899.