Loan Default Detection with Real-World Existing Imbalanced Financial Data

Sudha Senthilkumar[1*], Manas Vardhan[1], Patel Arya Mayurkumar [1], Brindha K[2]

School of Computer Science and Engineering , Vellore Institute of Technology, Vellore, India[1],
School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India[2]

[*] sudha.s@vit.ac.in

Abstract - Credit payments are sure a hassle-free way to live life but also introduce some degree of uncertainty with regards to repayments. More banks and non-banking financial institutions are focusing on methods to predict defaults effectively. Default detection makes a strong use case in Supervised Machine Learning. The aim of this research paper is to evaluate multiple Machine Learning models like SVM classifier,Decision Trees, Logistic Regression, Ensemble Method LightGBM and ANNs. Real-world data has been used in the training process of the above-said models implying it is severely imbalanced. Multiple sampling techniques have been used to train and evaluate the models like under-sampling, over-sampling, and SMOTE. Multiple metrics were used to evaluate the effectiveness of the models and a high degree of over-fitting was observed. AUC-ROC Score was used for evaluation to get a clearer picture.

Keywords: Default detection, Supervised Machine Learning, Sampling, SMOTE

I. INTRODUCTION

Credit Cards have made it easier for one to make transactions based on credit with the ability to pay later, or in easy monthly installments. With the emergence of easier and hassle-free credit options and an increase in the number of non-banking financial institutions, these services are now also available to the common man. Common man, however, is susceptible to making impulsive decisions.[8][9]

As of January of 2020, before the pandemic, a study by Slickdeals, a crowd-sourced shopping platform, showed that an Americans spent $155.03 on impulse purchases on average each month. In another poll conducted in April, during the pandemic, there was an increase in that number by 18% as itjumped to $182.98. [10][11]

This research aims to evaluate multiple classifiers like SVM, Logistic Regression, Ensemble Methods like LightGBM, Decision Trees and ANNs to obtain a classifier that performs the best at the binary classification of defaulted cases.[4-6]

Forecasting of financial time series has been proven to be effective using these models [3].

The data comes from reliable agencies like the Credit Bureau and contains the credit history of both defaulters and non-defaulters making our task a binary classification problem. Due to the nature of the data, it is imperative that the data will contain empty column entries and would have a class imbalance as there are more non-defaulters in the world than defaulters.[14][15]

Hence, the data would need to be handled to mitigate these problems before modeling. In addition to improving loan evaluation based on credit history, our research will be of great value to banks and NBFCs.

II. LITERATURE REVIEW

It is biased to favor the majority class when data are class-imbalanced. Under-sampling and oversampling can attenuate the problem by producing class-balanced data. SMOTE has been investigated both theoretically and empirically in this paper using simulated and real high- dimensional data.[1]

Synthetic samples are generated through the SMOTE method, which can help mitigate class-imbalance issues. This study shows that SMOTE is not as effective on high-dimensional data, particularly in the cases where the signal-to-noise ratio is low, for the classifiers considered. [2]

[3] For regression and classification issues, Extremely Random Trees (ET) offers a tree-based ensemble method. ET trains the tree using the whole training set, instead of using bagging as in Random Forest. The main benefit of ET algorithm is the computation efficiency and its robustness. This algorithm is used as the basis for our proposed method.

[4] The article provided a method based on k-means SMOTE and back propagation neural networks. It tries to provide a solution for unbalanced financial information related to credit card data. The SMOTE algorithm improved version helps to solve this issue.

III. DATASETS

The datasets such as application_train.csv, bureau.csv, bureau_balance.csv, POS_CASH_balance.csv, credit_card_balance.csv, previous_application.csv, installments_payments.csv, HomeCredit_columns_description.csv are used in this work.

IV. METHODOLOGY

In the first phase, data was cleaned and multiple data engineering techniques were used. The data contained a high number of empty values. Due to data being in abundance, empty rows were dropped.

After cleaning the data, it was checked for class imbalance. This class imbalance was handled by using two techniques: Under-sampling and SMOTE (Synthetic Minority Oversampling Technique. [14]

The classifiers trained on these training sets were SVM, Decision Trees, Logistic Regression, LightGBM, ANN

The neural network had the following architecture: Dense(100,activation='relu'),Dropout(0.2), Dense(50,activation='relu'),Dense(1,activation='sigmoid')

V. RESULTS

A.Accuracy score

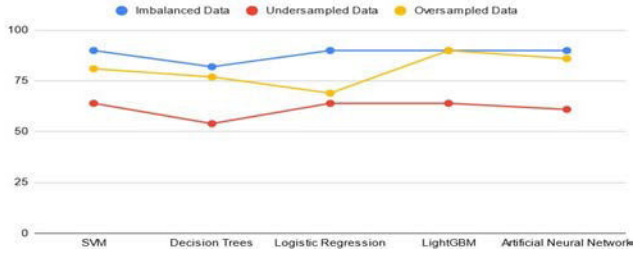

Fig. 1 Accuracy of the classifiers

B.AUC-ROC curve

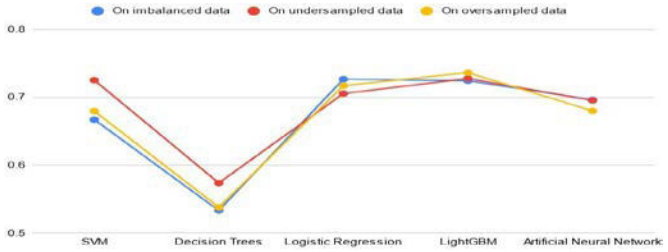

Fig. 2 AUC-ROC curve

The Light GBM classifier did a better job, it can not be ruled out that the classifier output was nowhere close to good by today's standards and business expectations.

VI. CONCLUSION

All of these modeling techniques are far from being perfect, as they; evidently, provide us with only half of the picture, giving us a weak estimate of a borrower's ability to repay his/her loan. However, more models like ARIMA can be used for financial data modeling they are reliable and are actively used by financial analysts all over the world to model and analyze the market and predict stock price patterns.

References

[1] D. A. Harahap, K. F. Ferine, N. Irawati, N. Nurlaila, & D. Amanah,Emerging advances in E-commerce: Panic and impulse buying during the COVID-19 pandemic, 2021.

[2] B. Rok, & L. Lusa, SMOTE for high-dimensional class- imbalanced data. BMC Bioinformatics, 14(1), 106-121, 2013.

[3] L. Zhu, D. Qiu, D. Ergu, C. Ying, & K. Liu, A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503-513, 2019.

[4] S. Rabiul Islam, W. Eberle, & S. Khaled Ghafoor, Credit Default Mining Using Combined Machine Learning and Heuristic Approach. arXiv e-prints, arXiv-1807, 2018.

[5] Y. Chen, & R. Zhang, Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network. Complexity, 2021.

[6] I. O. Eweoya, Adebiyi, A. A., Azeta, A. A., & Amosu, O., Fraud prediction in loan default using support vector machine. In Journal of Physics: Conference Series (Vol. 1299, No. 1, p. 012039). IOP Publishing., Aug 2019.

[7] M. Madaan, A. Kumar., C. Keshri, R. Jain, & P. 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.

[8] Y. Çelik, M. F. Aslan, & K. Sabancı, CLASSIFICATION (OF MAMMOGRAMS USING GLCM BASED) FEATURE EXTRACTION IN ARTIFICIAL NEURAL, NETWORK. In Academic Conference on Robotization, Engineering and Artificial Intelligence, 2019.

[9] U. Aslam, H. I. Tariq Aziz, A. Sohail, & N. K. Batcha, An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience, 16(8), 3483-3488, 2019.

[10] P. Supriya, M. Pavani, N. Saisushma, N. V. Kumari, & K. Vikas, Loan prediction by using machine learning models. International Journal of Engineering and Techniques, 5(22), 144-148, 2019.

[11] R. K. Amin, & Y. Sibaroni, Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region). In 2015 3rd International Conference on Information and Communication Technology (ICoICT) (pp. 75-80). IEEE, May 2015.

[12] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif, & S Islam, Application of machine learning in credit risk assessment: a prelude to smart banking. In TENCON 2019- 2019 IEEE Region 10 Conference (TENCON) (pp. 2023-2028). IEEE, Oct 2019.

[13] A. J. Hamid, & T. M. Ahmed, Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal, 3(1), 1-9, 2016.

[14] M. Vojtek, & E. Koèenda, Credit-scoring methods. Czech Journal of Economics and Finance (Finance a uver), 56(3-4), 152- 167, 2006.

[15] K. Alshouiliy, A. AlGhamdi, & D. P. Agrawal, AzureML based analysis and prediction loan borrowers creditworthy. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 302-306). IEEE. Mar 2020.