2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)

A Decision Tree Classifier based Ensemble approach to credit score classification

Ashok Maurya
Department of Economics,
School of Liberal Education,
Galgotias University, India
ashok.maurya@galgotiasuniversity.edu.in

Shivam Gaur
Department of Economics,
School of Liberal Education,
Galgotias University, India
shivamgaur9990@gmail.com

2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 979-8-3503-0611-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCIS60361.2023.10425039

***Abstract*— The process of classifying credit scores holds a crucial role in evaluating an individual's creditworthiness, influencing significant financial choices. This study is driven by the dynamic nature of credit scores and the financial sector's need for precise, real-time credit evaluations. This research introduces an ensemble-based method for credit score classification, utilizing a blend of diverse machine learning algorithms to improve accuracy and resilience. The ensemble approach capitalizes on each base classifier's strengths, mitigating biases, reducing overfitting, and enhancing overall classification accuracy. A comparison between the proposed model and existing frameworks demonstrates its competitive edge, surpassing many counterparts with an accuracy of approximately 92.25%. However, the study acknowledges the potential for further enhancement and validation across various datasets. The ensemble-based framework offers a promising avenue to heighten credit score classification accuracy, thereby contributing to informed financial decision-making and reinforcing credit ecosystem stability. Future endeavors involve expanding the model to include more datasets and refining data preprocessing techniques to achieve even more precise predictions.**

Keywords*— Ensemble approach, Bagging Classifier, Extra Tree Classifier, Random Forest, Stacking Classifier, Credit score classification

I. INTRODUCTION

A credit score is a numerical representation of an individual's or entity's creditworthiness, reflecting their ability to manage and repay borrowed funds. It serves as a standardized assessment tool used by financial institutions, lenders, and creditors to evaluate the risk associated with extending credit, such as loans, credit cards, mortgages, or other financial services. Credit scores [1] hold paramount importance in various financial contexts, shaping crucial decisions and outcomes. It is a linchpin in lending determinations, empowering lenders to make well-informed choices. A heightened credit score signifies diminished credit risk, indicating a greater likelihood of punctual repayment, thereby facilitating favorable lending terms encompassing reduced interest rates and augmented credit limits. Moreover, credit scores wield substantial influence over interest rate calculations for loans and credit products, bestowing preferential rates upon individuals with elevated creditworthiness. Notably, credit scores stand as pivotal gatekeepers for loan approvals, acting as a pivotal criterion for lenders to evaluate the feasibility of repayment. In a distinctive trajectory, credit scores extend their dominion to rental applications, enabling landlords to gauge the fiscal solidity of prospective tenants. Propitious credit scores can tilt the scales favorably for securing rental accommodations. Beyond lending and housing, credit scores infiltrate insurance dynamics, swaying insurance premiums as insurers factor credit scores into their pricing mechanisms. Enhanced credit scores correlate with reduced insurance costs. In this intricate interplay between finance, risk assessment, and opportunity, credit scores emerge as a multifaceted instrument influencing a spectrum of economic pursuits [2]. Credit scores are not static; they can change over time based on an individual's financial behavior. Responsible financial practices, such as paying bills on time and maintaining a low credit utilization ratio, can positively impact credit scores. Conversely, late payments, high debt levels, and other negative behaviors can lead to a decrease in credit scores. As such being so volatile it becomes imperative that proper frameworks must be deployed which would help classify credit score in real-time, thus facilitating the need for development of advanced credit classification techniques.

Credit score classification is a cornerstone of modern finance, providing an objective assessment of an individual's creditworthiness and shaping lending decisions that underpin economic activities. This multifaceted process involves analyzing vast troves of data, encompassing financial behaviors, demographic information, and historical transaction records, to derive predictive insights. The symbiotic fusion of data science methodologies and financial acumen has propelled credit score classification to new heights, enabling financial institutions to better gauge risk, allocate resources, and catalyze economic growth. This comprehensive exploration delves into the technical and economic dimensions of credit score classification, delving into its essence, the imperatives behind its evolution, and its myriad use cases in fostering a robust financial ecosystem. The complexity of contemporary financial systems and the diversity of individual financial profiles have propelled the need for sophisticated tools that distill intricate data into actionable insights. Credit score classification serves as a bridge between voluminous data and informed lending decisions. Its role becomes especially pivotal when lenders must assess the creditworthiness of applicants swiftly, accurately, and consistently. Human decision-making, while valuable, often succumbs to biases and subjectivity. In contrast, credit score classification offers a standardized, data-driven mechanism that enhances objectivity and reduces inherent biases. By integrating a variety of financial parameters, it empowers financial institutions to evaluate the likelihood of default and tailor lending terms more equitably.

The paper proposes a ensemble framework which combines the predictions by capitalizing on the model diversity of the various based models used as well as proving to be more robust being less prone to overfitting, reducing bias and variance [3-5]. The paper thus has been arranged as section 1 providing an introduction to the topic, section 2 highlighting the various past works and literatures present. Section 3 explaining the working of the model and the various algorithms involved, while section 4 presents the results of the model simulation and how it has performed, with an apt conclusion being mentioned in section 5.

II. LITERTATURE REVIEW

There have been various contributions in this field of the classification of credit score with the help of various machine learning algorithms. In 2019, [6] created a hybrid credit scoring model which made use of various ensemble array of feature selection methods like quadratic discriminant analysis (QDA), multilayer feed forward neural network (MLFN) to achieve an accuracy over of 87.05%. [7] proposed a model which made use of the Binary BAT Algorithm (BBA) to make optimized predictions, the model was tested on four different datasets, giving an accuracy of about 84.2%. With the implementation of data mining models in tandem with feature selection, [8] used Generalized Linear model (GLM) coupled with decision trees (DT) to make predictions, their model gave an accuracy of 87.68% with around 10.74% Type 1 error and 1.57% of Type 2 error. Using the stalog German credit data, [9] proposed a Deep Genetic Hierarchical Network of Learners (DGNHL), which was a 29 layer structure providing a prediction accuracy of about 94.60%, considered to be the standard among the existing literature. While comparing the various machine learning models like, Support Vector Machine, Random Forest, Bayesian Naïve Bayes, [10] concluded that Random Forest had the highest accuracy of 93%. Making use of gradient boosting approaches [11], proposed a framework which was tested on the German and Australian datasets. The model provided an accuracy of about 81% for the German and 88.4% for the Australian datasets.

From the above literature it could be concluded that there have been various models but none of them breached the accuracy mark of about 93% as a whole. The paper proposes a comprehensive model which leverage on the various ensemble models to stack up the results and provide better and accurate predictions, which would be more easily implementable as compared to the counterparts of DGHNL, BBA, GLM etc. The proposed study will make use of ensemble approach to average out the predictions of the base classifiers.

III. METHODOLOGY

To achieve a high accuracy as well as make the predictions more reliable, the framework is made to be an ensemble of various base classifiers stacked up against a final estimator which makes sense of the prediction of each unit and thus makes the final prediction. In order to achieve such a working, certain steps were followed, which can be found in the subsequent sections.

A. Data set

The basis of the whole framework is the achievement of high accuracy and efficacy, which was tested on the "Credit Score Classification Dataset". The said dataset is easily accessible on Kaggle network, proving to be reliable source of data for analysis and research purposes. The dataset is designed to predict the creditworthiness of a person based on various demographic and financial behaviors. The target class of this dataset encompasses three distinct values, Good, Standard and Poor, which label the credit score of a person depending on the various input parameters (Table 1). While in number there are 26 distinct feature of a particular person, which corroborate to about 50,000 various unique values. The various input parameters of a particular person are recorded over a period of four distinctive months. Such expansive inputs as well as size of data makes this dataset apt to test the framework and make relevant conclusions.

TABLE I. INPUT PARAMETERS

Parameter	Description
Annual_Income	Annual income in USD
Num_Bank_Acc	Number of bank accounts
Num_Credit	Number of credit cards held
Num_Loan	Number of loans taken
Delay_due	Number of days delayed from payment date
Delayed_payments	Number of delayed payments
Num_Credit_inq	Number of credit card inquiries
Out_Debt	Outstanding Debt in USD
CU_ratio	Credit utilization ratio
EMI_month	Total EMI per month
Month_bal	Represents monthly balance after EMIs
Credit_Age	Age of credit history

B. Classification Algorithms

Since the framework is an ensemble model the various algorithms used are as follows:

1) Bagging Classifier: The Bagging classifier, which stands for Bootstrap Aggregating classifier, functions as a binary classification algorithm rooted in an ensemble approach aimed at enhancing predictive accuracy [12,13]. This technique involves generating multiple subsets of the original training dataset through random sampling with replacement. Each of these subsets is then employed to train individual classifiers, typically built upon base classifiers like Decision Trees. The Bagging classifier amalgamates the outcomes of these individual classifiers by aggregating their predictions. By training multiple classifiers on diverse data subsets, the Bagging classifier diminishes the potential impact of singular classifier biases or noise in the data. The consolidation of predictions results in a more resilient and precise binary classification outcome.

2) Extra Trees Classifier: The ExtraTreeClassifier, short for Extremely Randomized Trees Classifier, is a binary classification algorithm, which constructs a collection of Decision Trees to predict whether an instance belongs to a certain class or not [14]. The ExtraTreeClassifier selects splitting points for nodes in a rather random manner. It doesn't meticulously search for the most optimal attribute to divide the data; instead, it makes these choices based on randomness. By introducing randomness into the decision-making process, the ExtraTreeClassifier becomes more robust to noise and variations in the data. This means it's less likely to overfit the training data, making it a suitable candidate for such an approach.

3) Random Forest: Random forest is yet another, powerful and versatile ensemble machine learning algortihm which works using the decision trees. The major difference of this method with the other two is the use of Bootstapped smapling of the data unlike the bagging classifier, by introducing diversity with feature selection [15]. The algorithm randomly selects subsets of the training data with replacement. This means that each subset might contain some repeated instances and miss others. Each subset is then used to train a separate Decision Tree. Such a robust and versatile architecure makes it fir to be used in such a scenario.

4) Histogram Gradient Boosting Classifier: It is a binary classification algorithm that belongs to the ensemble methods category [16]. This employs histogram-based data representations for guiding its decision-making process. Instead of analyzing individual instances, it groups data into bins and processes them in a more efficient manner. This methodology introduces a significant speed advantage in both training and prediction phases. By using histogram-based data representations, it captures important patterns while enhancing computational efficiency. It adapts a boosting framework, sequentially refining the model by correcting errors made by prior trees. Through the utilization of gradient descent and regularization techniques, the algorithm optimizes the predictive model's accuracy and stability.

5) XGBClassifier: Extreme Gradient Boosting Classifier (XGBClassifier) leverages an advanced gradient boosting framework that delivers remarkable predictive power. It sequentially builds an ensemble of Decision Trees, with each subsequent tree correcting the errors made by its predecessors. This iterative process ensures the creation of a highly accurate and robust predictive model [17,18]. A standout feature of the XGBClassifier is its focus on gradient descent optimization, which fine-tunes the model by minimizing the prediction errors. This results in a finely tuned predictive model that excels in handling complex relationships within the data.

6) Stacking Classifier: Stacking classifier works on the concept of combining multiple models to enhance the predicitve accuracy and robustness, generally built on top of many base classifiers and combine their predicitons to make a final one [19]. All the above base classifers which are ensemble in their nautre generally use a majority voting or aggreation but the differnce here is that this relies on the usage of a meta-classifier to combine predctions. The Stacking Classifier's unique feature lies in its ability to learn how to combine the strengths of different base classifiers. This often leads to improved predictive performance by leveraging the diverse expertise of individual base classifiers.

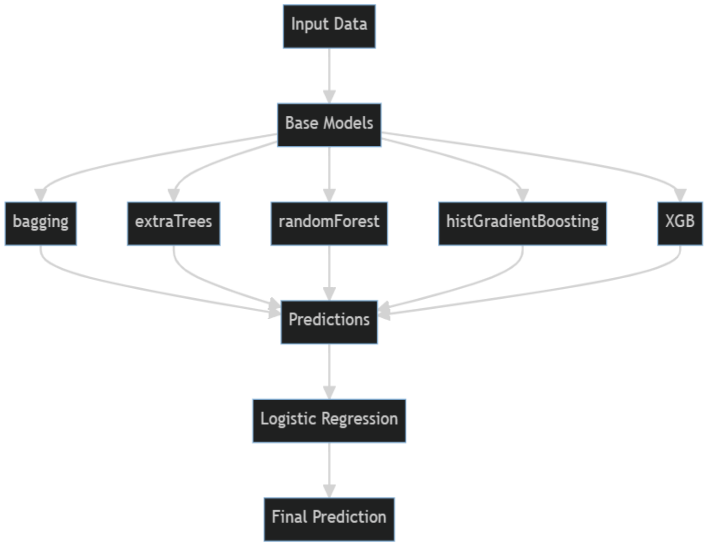

Figure 1: Working of proposed model

The main highlight of this working of the framework is that the meta-learner used is a logistic regression algorithm which makes sense of these sub-ensemble models and then makes the final prediction. The framework doubles (Figure 1) in by leveraging the variance reduction given by bagging classifier, enhanced diversity by extra tree classifier, optimized gradient descent by the XGBClassifier, the bias-variance tradeoff by Random Forest and histogram approach of Histogram Gradient Boosting Classifier and combines all these advantages to make a final and more reliable prediction.

C. Data Analysis

To better understand the type of the data that the model is being fed, exploratory data analysis was performed to understand the distribution of the target class (Figure 2) as well as the missing data in the various input parameters.

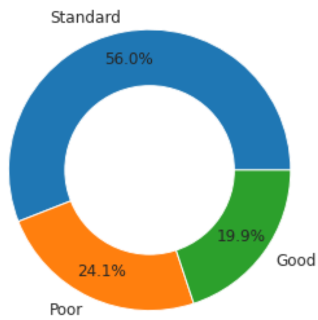

Figure 2: Credit Score Distribution

In continuation of exploring the data, it was found that much of the input parameters needed label encoding for them to be used in the decision-making process, while in figure 3 it can be seen that from an economic perspective the dataset is consistent with people having a good credit score have a higher annual income and a gradual shift leads to a decline in the credit score.

Figure 3: Annial Income and Credit Score

D. Data Preprocessing

The whole data set is composed of various financial aspects of an individual as such it is a combination of varied input parameters as such the various categorical data were label encoded to make the prediction easier. The data was also cleaned by the dropping of various unnecessary columns like name and occupation. Since it has already been established (Figure 1) that the target variable is highly imbalanced as such to tackle this issue the Synthetic Minority Oversampling Technique (SMOTE) [20] was used to generate samples of the target classes. Such steps lead to the data being more refined and appropriate to be used for further processing.

As the encoded dataset, had a wide range of values because of the annual income, number of credit card, monthly income columns, the better way to make them suitable for prediction was to scale them down with the help of the Power Transformer coupled with the Yeo-Johnson method. This transformation lead to the whole dataset being normalized and thus making it easier for the various algorithms to identify patterns and make predictions.

IV. Results and Discussion

The performance of the proposed model was gauged on the basis of various parameters of accuracy, precision, recall and f1 score, each highlighting the efficacy of the model as well as clears the air around the presence of false positives and false negatives. From table 2, it can be seen that the model performed pretty decent in terms of high accuracy, high precision proving to highly efficient. The table shows the test accuracy, while the train accuracy was about 99.89, proving the prowess of the algorithm in classifying credit score categories.

TABLE II. Model Parameters

Parameter	Score
Accuracy	92.2546
Precision	91.5962
Recall	93.4596
F1-score	92.3569

While comparing the model with the existing literature present, it can be seen that it has performed fairly better than the rest of the architecture present with the only exception being DGNHL, further details can be highlighted in figure 4. This can be used to highlight that the proposed model although being based on the relatively simpler algorithms could provide better results and thus score higher than most of the algorithms already seen. The advantage of the proposed model stands out from the rest based on the fact that the interpretability and ease of implementation of this model. The fact that the other models, require much computation power as well as are hard to implement with respect to hyperparameter tuning, feature engineering, scaling and robustness makes the ensemble model a fair easier and approachable method for credit score classification.

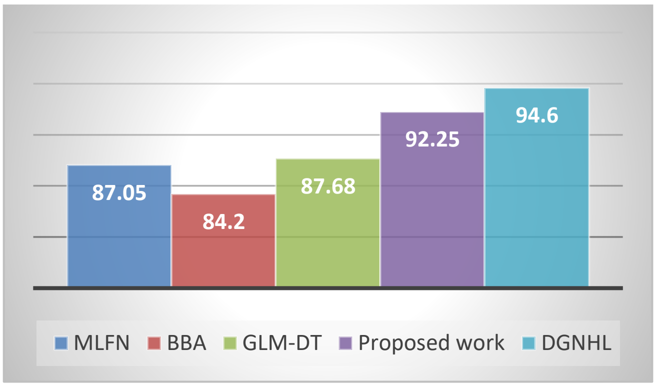

Figure 4: Comparision of various models (in terms of accuracy)

Conclusion

Through the study it can be seen that credit score is an integral part of the financial system and as such the correct classification is necessary to maintain the balance of the credit cycle. Previous works have shown an a high rate of accuracy in this field but failed short in many terms, owing to which this framework of an ensemble of ensemble models was proposed which makes use of the various base classifiers, their advantages like variance reduction, enhanced diversity, optimized gradient descent, the bias-variance trade off as well as the histogram data approach to build a model which shows an accuracy of 92.25% which is relatively higher than the industry standard being present. The model although robust, needs to be tested on other dataset to make it more reliable as well as there is call for better data preprocessing such that the model makes more accurate predictions.

References

[1] Arya, Shweta, Catherine Eckel, and Colin Wichman. "Anatomy of the credit score." *Journal of Economic Behavior & Organization* 95 (2013): 175-185.

[2] Chatterjee, Satyajit, et al. *A quantitative theory of the credit score*. No. w27671. National Bureau of Economic Research, 2020.

[3] Dong, Xibin, et al. "A survey on ensemble learning." *Frontiers of Computer Science* 14 (2020): 241-258.

[4] Dogan, Alican, and Derya Birant. "A weighted majority voting ensemble approach for classification." *2019 4th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2019.

[5] Dahiya, Manju, Naman Mishra, and Riya Singh. "Neural network based approach for Ethereum fraud detection." *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE, 2023.

[6] Tripathi, Diwakar, et al. "A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification." *Computational Intelligence* 35.2 (2019): 371-394.

[7] Tripathi, Diwakar, et al. "BAT algorithm based feature selection: Application in credit scoring." *Journal of Intelligent & Fuzzy Systems* 41.5 (2021): 5561-5570.

[8] Nalić, Jasmina, Goran Martinović, and Drago Žagar. "New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers." *Advanced Engineering Informatics* 45 (2020): 101130.

[9] Pławiak, Paweł, et al. "DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring." *Information Sciences* 516 (2020): 401-418.

[10] Trivedi, Shrawan Kumar. "A study on credit scoring modeling with different feature selection and machine learning approaches." *Technology in Society* 63 (2020): 101413.

[11] Lawi, Armin, Firman Aziz, and Syafruddin Syarif. "Ensemble GradientBoost for increasing classification accuracy of credit scoring." *2017 4th international conference on computer applications and information processing technology (CAIPT)*. IEEE, 2017.

[12] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." *Procedia computer science* 48.2015 (2015): 679-685.

[13] Sharaff, Aakanksha, and Harshil Gupta. "Extra-tree classifier with metaheuristics approach for email classification." *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*. Springer Singapore, 2019.

[14] Tamim Kashifi, Mohammad, and Irfan Ahmad. "Efficient histogram-based gradient boosting approach for accident severity prediction with multisource data." *Transportation research record* 2676.6 (2022): 236-258.

[15] Xuan, Shiyang, et al. "Random forest for credit card fraud detection." *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*. IEEE, 2018.

[16] Konar, Karabi, Saptarshi Das, and Samiran Das. "Employee attrition prediction for imbalanced data using genetic algorithm-based parameter optimization of XGB Classifier." *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*. IEEE, 2023.

[17] Aydin, Zeliha Ergul, and Zehra Kamisli Ozturk. "Performance analysis of XGBoost classifier with missing data." *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)* 2.02 (2021): 2021.

[18] Carmona, Pedro, Aladdin Dwekat, and Zeena Mardawi. "No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure." *Research in International Business and Finance* 61 (2022): 101649.

[19] Sesmero, M. Paz, Agapito I. Ledezma, and Araceli Sanchis. "Generating ensembles of heterogeneous classifiers using stacked generalization." *Wiley interdisciplinary reviews: data mining and knowledge discovery* 5.1 (2015): 21-34.

[20] Fernández, Alberto, et al. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." *Journal of artificial intelligence research* 61 (2018): 863-905.