ICIC Express Letters
Volume **18**, Number **1**, January **2024**

ICIC International ©2024 ISSN 1881-803X
pp. **27–37**

A LOAN DEFAULT PREDICTION MODEL USING MACHINE LEARNING AND FEATURE ENGINEERING

Suliman Mohamed Fati

College of Computers and Information Sciences
Prince Sultan University
66833 Rafha Street, Riyadh 11586, Saudi Arabia
sgaber@psu.edu.sa

Received February 2023; accepted April 2023

Abstract. *Bank loan default is one of the vital issues that may affect the banking sector. To avoid such an issue, the bank needs to analyze huge data, so machine learning (ML) is used to help in making accurate loan approval decisions. However, the presence of loan default is very minor in any dataset, which may lead to class imbalance and prediction bias. Another issue is the existence of the irrelevant variables that may cause prediction model overfitting. Thus, this study aims to overcome these two issues by incorporating the machine learning classifier with feature engineering and dataset resampling in order to produce accurate prediction. Hence, this study evaluates the performance of four machine learning classifiers, namely K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), on a public default loan dataset from lending club. After applying the data preprocessing, the proposed methodology used the feature engineering to eliminate the irrelevant features based on features correlation. Then, Adaptive Synthetic Sampling (ADASYN) was applied to managing the class imbalance issue. The experimental results showed the seriousness of the model overfitting issue as the four models performed better with the feature engineering and ADASYN as the accuracy enhanced significantly. Among the four models, augmented RF model outperformed the other models in terms of accuracy, precision, sensitivity, specificity, F1 score, and AUC with the values of 0.95, 0.97, 0.96, 0.8, 0.94, and 0.88, respectively.*

Keywords: Bank loan approval, Loan default, Machine learning algorithm, Prediction model, Class imbalance, Feature engineering

1. **Introduction.**

For any bank, financial loans are one of the crucial businesses that contribute to the bank's success. As the bank needs to ensure the loan will be given to the deserving people, the main concern for bank staff is to approve loans accurately to avoid any negative consequences [1,2]. Furthermore, any erroneous loan application validation may have a social and economic impact on the bank and customers [3]. One of the issues caused by improper loan application validation is loan default [4] whereby the debater fails to pay his/her debt obligations [3]. For conventional loan approval, a bank hires an expert to approve/disapprove the loan applications according to a formula known as the "5Cs", which stands for capacity, condition, character, collateral (ensuring loan repayment), and capital. A conventional method may be subject to subjectivity and inaccuracy because lenders may use their own personal experience and knowledge to assess [5]. Alternatively, 'Credit Score' method can be used to evaluate the loan applicant's eligibility based on the applicant's history of both credit and payment [6,7]. However, credit scoring requires both experts alongside statistical algorithms to determine the applicant's eligibility in the presence of huge information from different sources. Despite the fact that banks go through intensive inspection processes before approving a loan, this information is highly uncertain [1] even with the possibility to share such information among different banks.

DOI: 10.24507/icicel.18.01.27

With the emergence of Artificial Intelligence (AI), the loan status can be predicted in the presence of massive and noisy data, which can be retrieved from different entities (e.g., government, and other banks) [5,6]. However, such algorithms may suffer from two key issues [8]. The first one is the curse of dimensionality where there are so many input parameters needed to predict the loan status. The higher the dimensionality, the less accurate the prediction. Another critical issue is the fact that the presence of loan default is very minimal. This issue is defined as class imbalance whereby the classifier may take account of non-default cases at the expense of default cases. Such a tendency may make it difficult for the classifier to identify the default cases [49]. Therefore, this paper aims to investigate these two issues to improve prediction accuracy.

The contribution of this paper can be summarized as follows. First, an intensive review of quality papers was done to select the widely used machine learning algorithms in default loan detection. Second, we evaluated the performance of four machine learning algorithms on a generic dataset. Third, we developed an automated detection model by incorporating the feature engineering, to identify the most relevant features that will provide an accurate prediction, and synthetic data sampling to enhance the model performance on the concerned dataset.

The remainder of the paper is organized as follows. We discussed background and related works in Section 2, the dataset used and the implementation details in Section 3, the results in Section 4 and the conclusions in Section 5.

2. **Background and Related Work.** There are several studies in the literature that use data analysis in loan default prediction such as Artificial Neural Networks (ANNs) [11,12], fuzzy logic [13], and genetic algorithms [14,15]. On the other side, machine learning algorithms have the lion's share in the literature. Bae and Kim [16] and Arun et al. [17] discussed the importance of machine learning in assessing the eligibility of loan applicants in order to minimize credit risk factors. Turkson et al. [18] evaluated multiple ML algorithms (i.e., fifteen algorithms) models to produce the most accurate ML classifier in loan prediction. Based on their findings, LR provided the highest accuracy (81%) among its counterparts. The model selected three key features. Similarly, two other studies [19,20] used LR on uncorrelated features to predict loan approvals in different banking sectors. For instance, Vaidya [19] provided ML model wherein logistic regression model was used for loan approval prediction based on the applicant's record. Vaidya's model has the limitation of relying on independent estimation variables with large parameters' samples, despite LR's ability to oversee nonlinear data [21].

On the other hand, decision trees were used by [14,22,23]. According to [14], the credit rating for Iranian banks can be reduced by using Decision Trees (DT). On the other hand, Amin et al. [23] applied the C4.5 DT algorithm on a dataset of one thousand cases to predicting the loan status with a precision and accuracy of 96.4%. The Ensemble GradientBoost algorithm was also used by Lawi et al. [24] to identify loan defaulters based on a German-language dataset. The results showed an accuracy of 81%. Similarly, Yadav et al. [22] proposed similar work whereby the DT algorithm with k-fold was implemented to classify the loan applicants into multiple classes based on selective features. Priya et al. [25] presented another work whereby Random Forest (RF) was used to obtain an accuracy of 81%. Priya et al. [25] examined the loan features and concluded that credit history is the most critical feature in loan approval.

In contrast, based on realistic datasets obtained from the Bank of England, Taneja et al. [26] analyzed loan factors such as job status, weight, and income source using fuzzy rules. Taneja et al. [26] compared their works with different works and achieved accuracy of 83% on the given dataset. Jiang et al. [27] took a different approach where they analyzed the descriptive text of the loan to find soft features (e.g., part-of-speech features, sentiment features, and social relationship information).

Another way to use machine learning is to compare multiple machine learning algorithms to find the most appropriate algorithm based on the used data. Gahlaut et al. [28] developed a model to predict loan approval based on factors of occupation, financial status, and family circumstances. Gahlaut et al. used two algorithms named Random Forest (RF) and Linear Regression (LR). The results showed that RF outperforms LR with accuracy values of 0.79 and 0.76, respectively. Likewise, Arora and Kaur [29] tested four algorithms namely random forest, SVM, Naïve Bayes and K-Nearest Neighbors (KNN) to show how accurately they can predict loan defaulters. The Random Forest (RF) algorithm gave better accuracy than the other algorithms when used with optimized feature selection methods. Zhou et al. [30] introduced another work whereby five machine algorithms, namely random forest, decision tree, Bayes classification, Bagging, and Boosting, were assessed in the field of loan defaulter prediction. The results show that decision trees perform better. Although Zhou et al. [30] introduced a comprehensive explanation for the algorithms, the work lacks experimental results, in-depth analysis, or model evaluation. Alternatively, Arya et al. [9] conducted a comparison using a data-driven approach for three algorithms named Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM). The results showed that decision tree provided higher accuracy (72.05%) than the others.

Likely, Hamid and Ahmed [31] compared J48, Bayes Net, and Naïve Bayes algorithms. The highest accuracy that they achieve is 78%. Moreover, Turkson et al. [18] evaluated around fifteen machine learning models to find the most suitable algorithm to be used in loan prediction. Their experimental results showed that logistic regression provided the highest accuracy among its counterparts with 81% in a dataset with three selected features. Song and Peng [32] proposed a detailed evaluation framework for the loan classification models used by banks.

Likewise, Metawa et al. [15] compared different machine learning algorithms to predict mortgage defaulters. The algorithms used are Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Factorization Machines (FM). As per their study, FM showed better performance as a nonparametric algorithm. Furthermore, Kemalbay and Korkmazoğlu [20] applied LightGBM, XGBoost, random forest, and logistic regression algorithms to building a binary prediction model that predicts loan defaulters. Furthermore, a random forest model showed better performance with AUC values of 0.89 and accuracy values of 0.88. In addition, the proposed evaluation model is based on Multiple Criteria Decision Making (MCDM), in which the different models are evaluated against a set of performance metrics to determine the most suitable model. Such results were also presented in [33].

As discussed before, the default loan detection has been investigated in the literature with introducing many proposed works. In addition, the machine learning algorithms were deployed. Each work has its own approach to enhancing the performance. However, most of the cited works lack the ability to deal with both high dimensionality and class imbalance, which are critical issues in default loan [8]. Thus, this gap motivates us to investigate and try to resolve these two issues in this study aiming to improve the predictors' performance.

3. The Proposed System.

A primary goal of this research is to predict whether loan applicants are likely to default on their loans based on historical loan default data. The loan will be labeled as "default loan" or "non-default loan" based on the classification model that we use. The novelty of this work is twofold. First, our model eliminates the features with less impact on the prediction using the feature engineering to enhance the prediction accuracy. Second, our model reduces the bias to the class with high publication using Adaptive Synthetic Sampling (ADASYN) technique. Incorporating these two

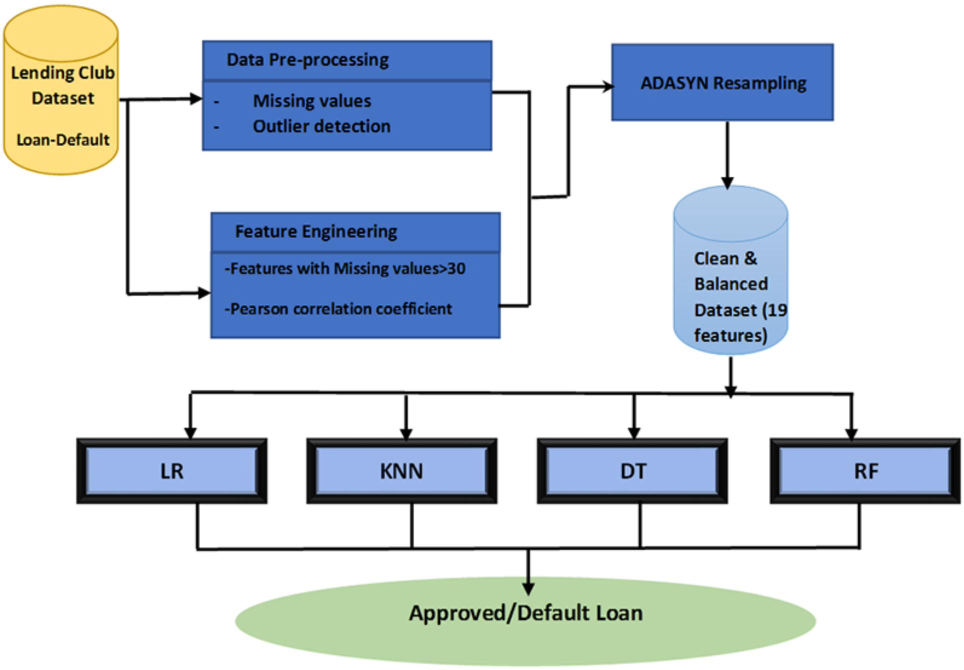

FIGURE 1. The proposed loan prediction methodology

techniques will enhance the prediction performance and accuracy. Thus, we adopted the methodology shown in Figure 1.

The first stage of the proposed methodology is data collection. In the collection stage, we have used a public dataset for loan default prediction by Lending Club [34] from Kaggle [40]. Following that, the necessary data preprocessing was applied to detecting both missing values and outliers. Next, the Pearson correlation coefficient was calculated to select the most relevant features. Adaptive synthetic sampling was applied to alleviating class imbalance. The clean dataset was then used for prediction through four machine learning algorithms, namely LR, KNN, DT, and RF algorithms. The performance of four predictors was evaluated using different metrics to monitor the model behavior.

3.1. **Dataset description and preprocessing.** In order to validate the proposed model and benchmark the results with other models, a loan default prediction dataset, which is a public dataset by Lending Club, was used. This collected data represents the loans for the period from 2007 to the 3rd quarter of 2017. The dataset consists of 1,345,310 records with 150 features [45]. The loan status variable has been encoded with 1 and 0, which indicate whether the loan is charged off or not, respectively. In this experiment, we focused on preventing the prediction model overfitting throughout two steps. The first step is to determine which features are relevant in the prediction process. As mentioned earlier, there are 150 features, so we need to apply feature extraction technique to selecting the most prominent features that produce accurate results. The second step is to check the class imbalance, which is critical in model overfitting. In this experiment, we found the 9% of the total records is labeled with default, which shows the ratio of default cases to the fully paid class imbalance in this dataset. The dataset was partitioned to training set (80%) and testing set (20%). The dataset was then loaded into the Google Colab notebook by mounting the drive.

3.2. **Feature engineering of default dataset.** In order to determine the influential features that contribute to the default prediction, we follow the following steps. First, we checked the missing values in the different 150 features to find if there are any missing values that may affect the prediction, so the heatmap was used. We analyze the missing values in the 150 features to see the percentage of the missing values in each feature. Then, we dropped the features that have more than 30% missing values. As per the heatmap analysis, we found that the majority of missing values exist in emp_title, emp_length, revol_util, and mort_acc features. In addition, the feature selection process was done in two stages. In the first step, thirty features with the strongest correlation to the target variable are filtered by Recursive Feature Elimination (RFE) to gradually eliminate them. We compute the correlation between the remaining features and the descriptive feature "charged_off". The following diagram depicts the correlation values in ascending order. As shown in Figure 2, last_pymnt_amnt, int_rate_fico_range_low, and fico_range_high have the highest correlation. However, all the features were retained with a correlation above 0.03 to ensure stronger association that may help in enhancing the prediction accuracy.

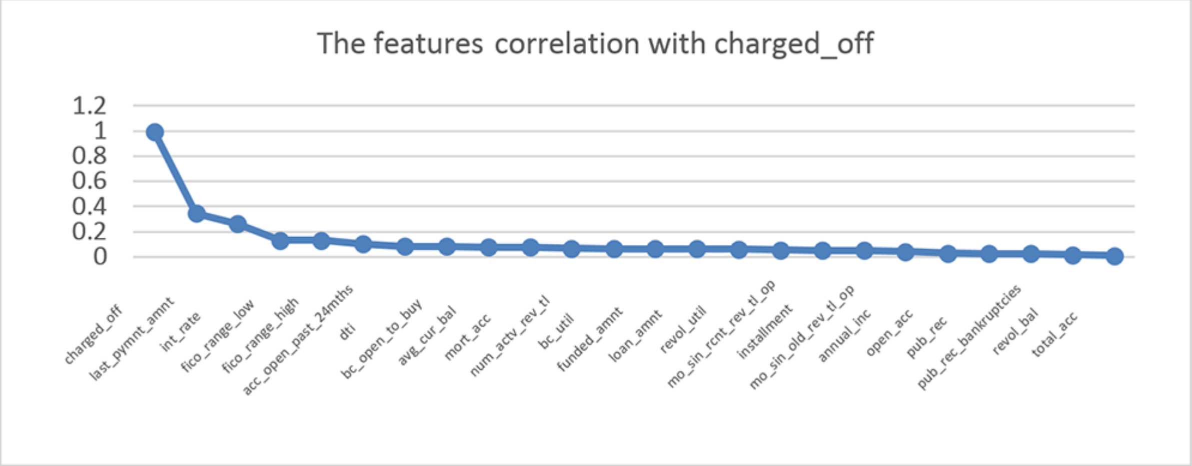

Figure 2. The presentation of feature correlation

Next, the redundant features are identified by bootstrapping the Pearson correlation coefficient between the charged_off values and the other features values. Nineteen features were finally filtered out. Equation (1) explains the correlation coefficient process.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{1}$$

where r is the correlation coefficient, x_i is the value of charged_off variable value, $\bar{x}$ is the mean of charged_off variables, y_i is the values of other features, and $\bar{y}$ is the mean of other features values.

To rank the selected features according to their importance, the feature importance was taken into account at the end of the feature selection process. In this step, the Random Forest (RF) algorithm is used to determine the feature importance [45,46]. After selecting the influential features that prevent the model from overfitting, it is time to fill in the missing values of the selected features. Missing values and outliers are common issues in the dataset that may affect the prediction accuracy. We aim to ensure the dataset is complete and harmonized. To achieve this goal, missing values and outliers were checked. To detect the outliers in the concerned dataset, we used a univariate method based on the Median Absolute Deviation (MAD) [48]. MAD is calculated based on a range around the median, multiplied by a constant, which varies according to the feature. As per our observation, there are four features with outliers, which are annual_inc, loan_amnt, term, and installment. To ensure better removal of outliers, the trimmed mean was applied with a subjective cutoff of 20% that disregards extreme values. Next, the mean was calculated for each feature and replacing the missing value.

3.3. **Class imbalance in loan default and resampling techniques.** One of the most critical issues in machine learning is class imbalance. This can lead to model overfitting due to noisy training data that negatively affects the model performance and causes it to produce incorrect decisions in testing. Such noise includes the class imbalance where the majority of data belong to one class, while the other class has very minor members. Thus, the model will be biased towards the majority class at the expense of the minor class. As we previously discussed, loan default dataset contains two classes: fully paid with 96% of the dataset, whereas only 4% of the dataset is chegred_off, as shown in Figure 3. Such diverse cause the class imbalance. Therefore, solving this imbalance issue is critical before running the prediction model.

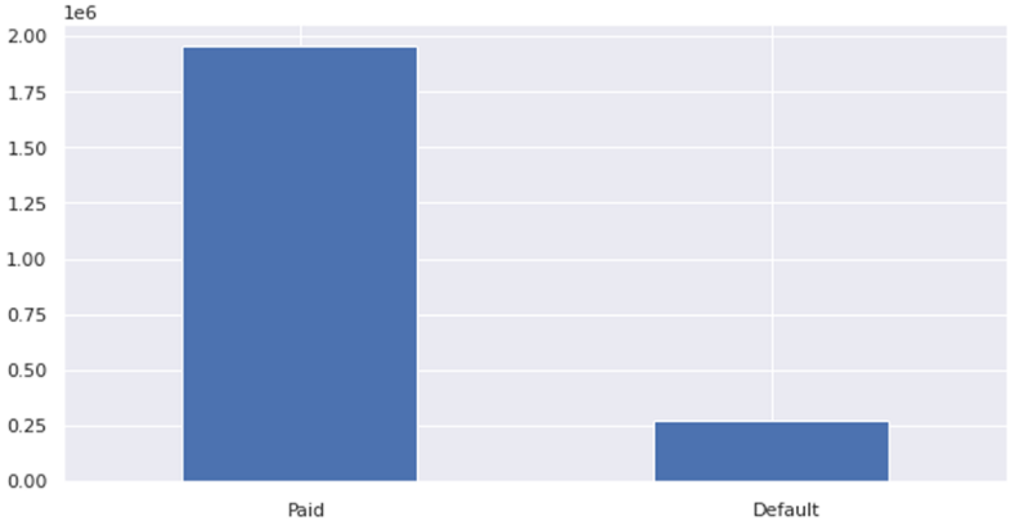

FIGURE 3. The class imbalance in loan default dataset

In order to overcome this issue, there are two mechanisms, namely under-sampling and oversampling. The under-sampling focuses on deleting some examples from the majority class to provide a compact balanced training set, while the oversampling focuses on duplicating some examples from the minority class. Under-sampling has the advantage of reducing the cost of learning. However, it has some limitations, such as increasing class variance and giving wrong predictions [41] or ignoring relevant examples that are necessary for the algorithm to learn [42]. Therefore, oversampling techniques aim to replicate the minority class instances to rebalance the dataset. SMOTE is one of the good examples for this category [43,44]. In the current work, Adaptive Synthetic Sampling (ADASYN) algorithm, which is a variant of SMOTE oversampling technique, was used. The essence of ADASYN is to use the density distribution of the dataset to decide the required synthetic samples to be produced for each minority data example. Then, ADASYN assigns different weights to the existing data examples in the minority class based on their learning difficulty level. Based on the k nearest neighbors of a sample, the percentage of samples to be synthesized for that sample relative to the total number of samples is calculated. This iterative process can continue with normalizing the distribution criteria until producing synthetic minority data that meets the distribution criteria. As ADASYN produces balanced presentations for the minority class, the machine learning process is forced to consider the difficulty to learn examples as well [48].

4. **The Proposed Prediction Module Results and Discussion.** After performing pre-processing, feature selection, and data imbalance, various classifiers have been used to predict whether the loan is in default or not. The classifier models constructed are K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). The four classifiers were evaluated on the loan default dataset, generated

by lending club. The confusion matrix, which is one of the most critical metrics to measure the classifier's performance, was generated for all the classifiers. The confusion matrix includes the true positive (actual legitimate loan) samples, true negative (actual default loan) samples, false positive (default loan identified as legitimate) samples, and false negative (legitimate loan identified as default) samples. In addition, the classifiers' performance was evaluated using accuracy, precision, sensitivity (aka Recall), and F1 score.

Besides accuracy, specificity, and sensitivity, Area Under Curve (AUC) is an influential performance metric that measures TPR (sensitivity) versus FPR (specificity) [35-38]. AUC and ROC (Receiver Operating Characteristics) show the performance of the classification algorithm at all classification thresholds [35,36]. They measure the performance of the model in different points while the model is operational [37,39].

Therefore, this section records the experimental results, according to the above metrics, of the proposed prediction model using the four machine learning algorithms (LR, KNN, DT, and RF), and then records the results with and without combining those algorithms with the resampling technique, ADASYN. In the first experiment, we ran the four ML algorithms on the dataset without treating the class imbalance. This is to monitor the predictors' behavior in the presence of class imbalance and the model overfitting. Figure 4 depicts the results for the four models, namely LR, KNN, DT, and RF. As shown in the figure, all the models have low accuracy 80% and lower. The KNN has the worst performance in terms of accuracy (65%), precision (76%), sensitivity (68%), specificity (57%), F1 score (68%), and AUC (0.63), while RF has the best performance in terms of accuracy (80%), precision (87%), sensitivity (87%), specificity (50%), F1 score (87%), and AUC (0.68). In the best cases, the proposed algorithms used can identify only 80% of default loan cases. Such value is unacceptable in the banking sector. As per the sensitivity analysis of algorithms used, the four models have less ability to detect the legitimate loan instances (low sensitivity), as well as less ability to detect the default loan instances (low specificity), which in turn, justify the low values of Area Under the Curve (AUC) values in all the algorithms (AUC is less than 0.7).

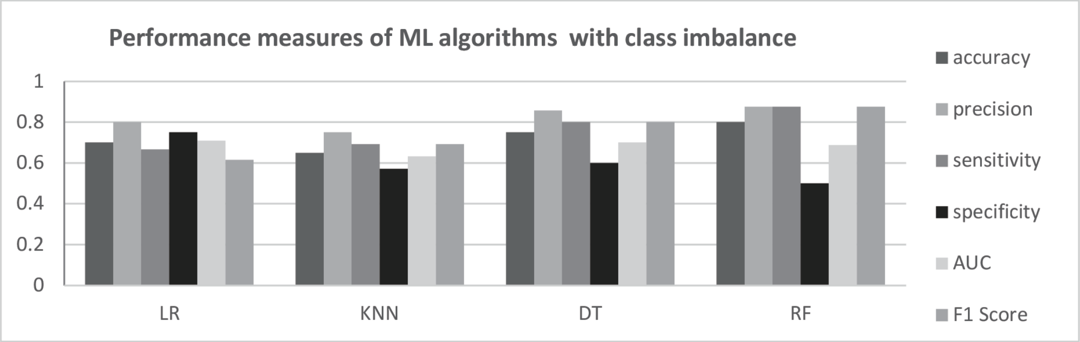

Figure 4. Performance measure of the models in the presence of class imbalance

Another experiment was conducted to show the performance of the four algorithms with the integration of the Adaptive Synthetic Sampling (ADASYN) technique. Figure 5 depicts the results of applying the resampling to the four algorithms.

As a comparison of Figures 4 and 5, applying resampling techniques, particularly ADASYN, improves the performance of all the algorithms. For instance, the accuracy of the LR, KNN, DT, and RF improved from 0.7, 0.65, 0.75, and 0.8 (without sampling) to 0.9, 0.87, 0.91, and 0.95 (with sampling). Such accuracy improvement can be justified by the ability of ADASYN to redistribute the minor class. Thus, the models would be able to avoid bias and model overfitting during the prediction process. As shown in Figure 5,

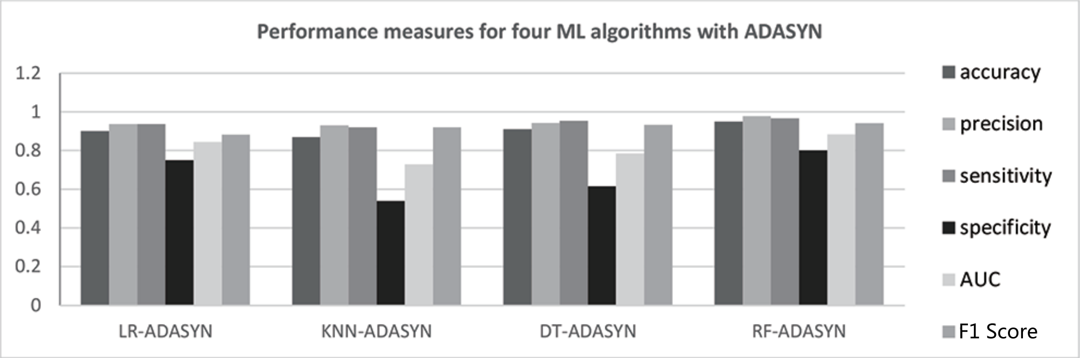

Figure 5. Performance measure of the augmented models with ADASYN

the models were able to improve their sensitivity and specificity. Both performance metrics show an enhanced ability to identify positive and negative cases correctly. Certainly, the values need improvement, and we intend to make improvements in the future. With accuracy of 0.95, precision of 0.97, sensitivity of 0.96, specificity of 0.8, F1 score of 0.94, and AUC value of 0.88, RF displayed superior performance. In addition, the performance variation of DT was significant in the experiment. Despite the close results of DT and RF in our experiments, RF outperforms DT. Without doubt, this superiority came from the idea of RF as a set of linked decision trees that work on the large dataset. Each decision tree comes with its own results, and then the results are aggregated based on the voting approach. With such voting, RF is able to select the best results and provide better performance. With this improvement, the banking industry will be able to predict loan status based on the most relevant features. ADASYN helps in ensuring the prediction will not omit the rare default loan case. Therefore, the learning process will be more active to allow the loan agency to detect the loan default with better performance.

Comparison with similar works. As per the literature review, there are different works in the field of loan prediction. However, we tried to find works that use the same dataset in order to provide realistic comparisons. There are two remarkably similar works proposed by Muneer and Fati [35], which utilized LightGBM, XGBoost, RF, and LR, and the other work proposed by Leys et al. [47], which used RF, DT, SVM, and LR. Both the current work and the other two use RF and LR in a similar way, but with the added advantage of ours in integrating feature reduction and ADASYN. Besides, as RF provides the best performance in our experiments, we feel these two studies are sufficient to compare with them. The comparison results are depicted in Figure 6 based on AUC value. The AUC value takes both sensitivity and specificity into account, so it is enough to determine which model is better.

The RF-ADASYN model provided the highest AUC value with 0.88, which means the highest prediction ability, followed by LightGBM [35], and RF [47]. This can be interpreted by the superiority of random forest with ADASYN to identify the default cases correctly. As explained earlier, RF consists of a set of decision trees with a voting system. This helps RF to explore more cases in parallel in order to come up with the highest accuracy based on the performance of all the dependent decision trees. Utilizing RF with ADASYN enabled the RF to run on synthetic datasets that take account of class balance. Such integration directs the RF to avoid any prediction bias, and provide the highest accuracy. Another notable point is the variation in the values between LR-ADASYN and the LR in both studies [35] and [47] where LR-ADASYN provided AUC value of 0.84375 significantly higher than both LR [35] value of 0.6 and LR [47] value of 0.73. This can be referred to the ability of our model to oversee model overfitting through both class imbalance and feature reduction.

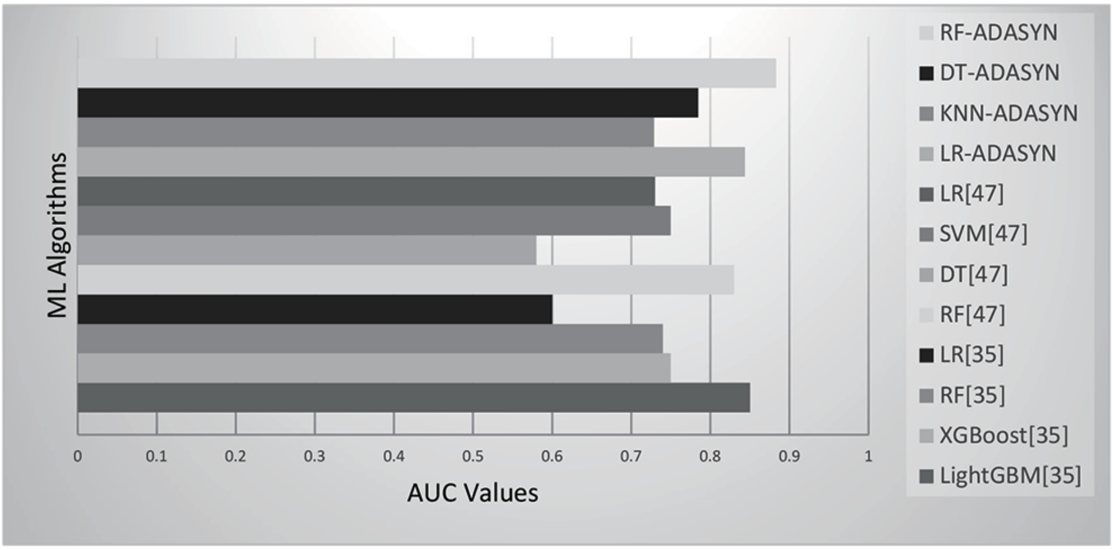

Figure 6. Comparison results with similar works

5. **Conclusion and Future Work.** Among the different techniques used by banks to detect the loan default, machine learning is widely used. However, two critical issues are associated with the use of machine learning in the presence of huge and noisy data, namely dimensionality curse and class imbalance. In this paper, different machine learning methods were investigated based on the literature review, which are KNN, LR, DT, and RF. Additionally, an augmented machine-learning-based prediction model was introduced by integrating Adaptive Synthetic Sampling (ADASYN) with the four ML models following feature engineering to identify the most influential features. Experimental results showed that the augmented Random Forest (RF-ADASYN) model outperforms similar approaches. Future research directions should focus on improving the model's performance through investigating more machine learning algorithms such as SVM, AdaBoost, and Gradient Boost. Besides, effective feature engineering and feature fusion techniques will be incorporated. In addition, enhancing ADASYN techniques by verifying the resultant samples and enhancing class balancing is another direction. An alternative approach is to build ensemble machine learning models by combining multiple machine learning algorithms and optimizing the algorithms' parameters to produce adaptive prediction models. For evaluation of the applicability of the model and its variations on real cases, it might be worthwhile to use real datasets from local entities.

REFERENCES

[1] S. M. Fati, Machine learning-based prediction model for loan status approval, *Journal of Hunan University Natural Sciences*, vol.48, no.10, pp.1-8, 2021.

[2] A. Gupta, V. Pant, S. Kumar and P. K. Bansal, Bank loan prediction system using machine learning, *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, pp.423-426, DOI: 10.1109/SMART50582.2020.9336801, 2020.

[3] K. Kohv and O. Lukason, What best predicts corporate bank loan defaults? An analysis of three different variable domains, *Risks*, vol.9, no.2, pp.29-39, DOI: 10.3390/risks9020029, 2021.

[4] A. S. Aphale and S. R. Shinde, Predict loan approval in banking system machine learning approach for cooperative banks loan approval, *International Journal of Engineering Research & Technology*, vol.9, no.8, pp.991-995, 2020.

[5] M. Madaan, A. Kumar, C. Keshri, R. Jain and P. Nagrath, Loan default prediction using decision trees and random forest: A comparative study, *IOP Conference Series: Materials Science and Engineering*, vol.1022, no.1, pp.12-42, 2021.

[6] M. I. Ahmed and P. R. Rajaleximi, An empirical study on credit scoring and credit scorecard for financial institutions, *Int. Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol.8, no.1, pp.275-279, 2019.

[7] S. V. Kulkarni and S. N. Dhage, Advanced credit score calculation using social media and machine learning, *Journal of Intelligent & Fuzzy Systems*, vol.36, no.3, pp.2373-2380, 2019.

[8] N. S. Alfaiz and S. M. Fati, Enhanced credit card fraud detection model using machine learning, *Electronics*, vol.11, no.4, pp.662-684, 2022.

[9] S. Arya, C. Eckel and C. Wichman, Anatomy of the credit score, *Journal of Economic Behavior & Organization*, vol.95, no.1, pp.175-185, 2013.

[10] D. Tripathi, D. R. Edla, A. Bablani, A. K. Shukla and B. R. Reddy, Experimental analysis of machine learning methods for credit score classification, *Progress in Artificial Intelligence*, vol.10, no.3, pp.217-243, 2021.

[11] B. Yang, L. X. Li, H. Ji and J. Xu, An early warning system for loan risk assessment using artificial neural networks, *Knowledge-Based Systems*, vol.14, no.5, pp.303-306, 2001.

[12] A. K. I. Hassan and A. Abraham, Modeling consumer loan default prediction using ensemble neural networks, *2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE)*, Khartoum, Sudan, pp.719-724, DOI: 10.1109/ICCEEE.2013.6634029, 2013.

[13] B. Kumar, I. Bawane, A. Shirsathe and P. Pardeshi, An expert system based on fuzzy logic for automated decision making for loan approval, *International Journal of Recent Advances in Multidisciplinary Research*, vol.2, no.12, pp.1078-1082, 2016.

[14] Z. Somayyeh and M. Abdolkarim, Natural customer ranking of banks in terms of credit risk by using data mining a case study: Branches of Mellat Bank of Iran, *Jurnal UMP Social Sciences and Technology Management*, vol.3, no.2, pp.307-316, 2015.

[15] N. Metawa, M. K. Hassan and M. Elhoseny, Genetic algorithm based model for optimizing bank lending decisions, *Expert Systems with Applications*, vol.80, no.1, pp.75-82, 2017.

[16] J. K. Bae and J. Kim, A personal credit rating prediction model using data mining in smart ubiquitous environments, *International Journal of Distributed Sensor Networks*, vol.11, no.9, pp.1-6, 2015.

[17] K. Arun, G. Ishan and K. Sanmeet, Loan approval prediction based on machine learning approach, *IOSR Journal of Computer Engineering*, vol.18, no.3, pp.18-21, 2016.

[18] R. E. Turkson, E. Y. Baagyere and G. E. Wenya, A machine learning approach for predicting bank credit worthiness, *2016 3rd International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, Lodz, Poland, pp.1-7, DOI: 10.1109/ICAIPR.2016.7585216, 2016.

[19] A. Vaidya, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, DOI: 10.1109/ICCCNT.2017.8203946, 2017.

[20] G. Kemalbay and O. B. Korkmazoğlu, Categorical principal component logistic regression: A case study for housing loan approval, *Procedia – Social and Behavioral Sciences*, vol.109, no.1, pp.730-736, 2014.

[21] A. Khan, E. Bhadola, A. Kumar and N. Singh, Loan approval prediction model a comparative analysis, *Advances and Applications in Mathematical Sciences*, vol.20, no.3, pp.427-435, 2021.

[22] O. Yadav, C. Soni, S. Kandakatla and S. Sawant, Loan prediction system using decision tree, *International Journal of Information and Computing Science*, vol.6, no.5, pp.137-143, 2019.

[23] R. K. Amin, Indwiarti and 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), *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, Nusa Dua, Bali, Indonesia, pp.75-80, DOI: 10.1109/ICoICT.2015.7231400, 2015.

[24] A. Lawi, F. Aziz and S. Syarif, Ensemble GradientBoost for increasing classification accuracy of credit scoring, *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, Kuta Bali, Indonesia, pp.1-4, DOI: 10.1109/CAIPT.2017.8320700, 2017.

[25] K. U. Priya, S. Pushpa, K. Kalaivani and A. Sartiha, Exploratory analysis on prediction of loan privilege for customers using random forest, *International Journal of Engineering Technology*, vol.7, no.2, pp.339-341, 2018.

[26] S. Taneja, B. Suri, S. Gupta, H. Narwal, A. Jain et al., A fuzzy logic based approach for data classification, *Data Engineering and Intelligent Computing*, vol.542, pp.605-616, 2018.

[27] C. Jiang, Z. Wang, R. Wang and Y. Ding, Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending, *Annals of Operations Research*, vol.266, no.1, pp.511-529, 2018.

[28] A. Gahlaut, Tushar and P. K. Singh, Prediction analysis of risky credit using data mining classification models, *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, DOI: 10.1109/ICCCNT.2017.8203982, 2017.

[29] N. Arora and P. D. Kaur, A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment, *Applied Soft Computing*, vol.86, no.1, 105936, 2020.

[30] P. Y. Zhou, K. C. Chan and C. X. Ou, Corporate communication network and stock price movements: Insights from data mining, *IEEE Transactions on Computational Social Systems*, vol.5, no.2, pp.391-402, 2018.

[31] A. J. Hamid and T. M. Ahmed, Developing prediction model of loan risk in banks using data mining, *International Journal on Machine Learning and Applications*, vol.3, no.1, pp.1-9, 2016.

[32] Y. Song and Y. Peng, A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction, *IEEE Access*, vol.7, pp.84897-84906, 2019.

[33] A. Bagherpour, *Predicting Mortgage Loan Default with Machine Learning Methods*, Ph.D. Thesis, University of California, USA, 2017.

[34] A. Coşer, M. M. Maer-Matei and C. Albu, Predictive models for loan default risk assessment, *Economic Computation & Economic Cybernetics Studies & Research*, vol.53, no.1, pp.149-165, 2019.

[35] A. Muneer and S. M. Fati, A comparative analysis of machine learning techniques for cyberbullying detection on Twitter, *Future Internet*, vol.12, no.11, pp.187-200, 2020.

[36] Google Developers, *Classification: ROC Curve and AUC*, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, 2021.

[37] B. K. Padhi, S. Chakravarty and B. N. Biswal, Anonymized credit card transaction using machine learning techniques, *Advances in Intelligent Computing and Communication*, pp.413-423, 2020.

[38] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, Real-time credit card fraud detection using machine learning, *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, pp.488-493, 2019.

[39] D. Sarkar, R. Bali and T. Sharma, *Practical Machine Learning with Python. A Problem-Solvers Guide to Building Real-World Intelligent Systems*, 1st Edition, Berkely, Apress, 2018.

[40] Kaggle, *Deep Learning & Analysis: Loan Default Prediction*, https://www.kaggle.com/slythe/deep-learning-analysis-loan-default-prediction, 2021.

[41] A. Dal Pozzolo, O. Caelen and G. Bontempi, When is undersampling effective in unbalanced classification tasks?, in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science*, A. Appice, P. Rodrigues, V. S. Costa, C. Soares, J. Gama and A. Jorge (eds.), Cham, Springer, 2015.

[42] M. Wasikowski and X.-W. Chen, Combating the small sample class imbalance problem using feature selection, *IEEE Transactions on Knowledge and Data Engineering*, vol.22, no.10, pp.1388-1400, 2009.

[43] V. García, R. A. Mollineda and J. S. Sánchez, On the k-NN performance in a challenging scenario of imbalance and overlapping, *Pattern Analysis and Applications*, vol.11, no.3, pp.269-280, 2008.

[44] D. A. Cieslak and N. V. Chawla, Start globally, optimize locally, predict globally: Improving performance on imbalanced data, *2008 8th IEEE International Conference on Data Mining*, Pisa, Italy, pp.143-152, DOI: 10.1109/ICDM.2008.87, 2008.

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

[46] X. Li, D. Ergu, D. Zhang, D. Qiu, Y. Cai and B. Ma, Prediction of loan default based on multi-model fusion, *Procedia Computer Science*, vol.199, pp.757-764, 2022.

[47] C. Leys, M. Delacre, Y. L. Mora, D. Lakens and C. Ley, How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration, *International Review of Social Psychology*, vol.32, no.1, 2019.

[48] H. He, Y. Bai, E. A. Garcia and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, pp.1322-1328, 2008.

[49] Z. Qing, Q. Zeng, H. Wang, Y. Liu, T. Xiong and S. Zhang, ADASYN-LOF algorithm for imbalanced tornado samples, *Atmosphere*, vol.13, no.4, 544, DOI: 10.3390/atmos13040544, 2022.