Bank Loan Default Prediction Model Using Machine Learning: A Comparative Study

Suliman Mohamed Fati [1†],

College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia

Summary
One of the issues caused by the improper loan application validation is loan default whereby the debater will not be able to pay his/her financial obligations incurred. To avoid such issues, the bank needs to analyze huge data to come up with a proper decision. Therefore, Machine learning is a promising direction to give accurate and on-time decision to predict the loan defaulter. The aim of this paper is to minimize the credit risk by predicting the loan default based on different loan factors or features. First, the collected data will be cleaned using different preprocessing techniques. Next, the most influential features will be identified using the correlation between the features. Once the data preprocessing done, four machine learning algorithms will be trained and tested, which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The novelty of this research can be represented in analyzing the behavior of four machine learning algorithms with different resampling techniques to find the most accurate prediction model. The experimental results showed the superiority of combination of Logistic Regression and over/under sampling on the other three algorithms in terms of accuracy precision, Recall, F1, and Area under curve (AUC). Such findings demonstrated the ability of Logistic Regression with resampling to predict the loan default based on the historical data provided..

Keywords:
Loan Approval, Machine Learning Algorithm, Logistic Regression, Loan Defaulter, Prediction Model, Random Forest.

1. Introduction

Financial Loans are one of the critical businesses for any bank that contribute to the banks' success. The bank loan has a two-side effect with both profit opportunities and risks. Certainly, the bank aims to increase the profit opportunities, while minimizing the negative risks. Therefore, the main concern for bank staff is to approve the loans for only the deserving people to avoid any negative consequences [1]. Consequently, the loan checking and validation should undergo rigorous validation process. Without any doubt, this validation process should be done in an acceptable time limit with no delay affecting the bank staff and applicants [2]. Furthermore, any erroneous loan application validation may lead to negative social and financial impact on bank, employees, clients, and other stakeholders [3]. One of the issues caused by the improper loan application validation is loan default [4]. The loan default is the inability of the debater to pay his/her financial obligations incurred [3]. Besides, among all the side effects of COVID-19 pandemic, the individuals' financial position is hugely affected, which in turn increased the rate of loan defaulter. All these varying and diverse information with high uncertainty will affect the loan process with no surety that the approved applicant will be good payer or defaulter [1]. Even through banks run an intensive validation and verification mechanism for the approval of a loan with scoring the loan risks. As one of the options to overcome this issue, the banks may collaborate to share the needed information that help in deciding the borrower's eligibility for a loan [1]. To achieve this goal, collecting and analyzing the historical data of customers from different financial entities is a good approach to get accurate results to minimize the probability of loan defaulter. To facilitate such information sharing, the governments may initiate credit bureaus in order to allow the different banks share the credit information to get accurate and better loan approval decisions. For instance, Saudi Credit Bureau (Simah) is the first and sole licensed national credit bureau offering consumer and commercial credit information services to respective members in the Kingdom of Saudi Arabia. Consequently, the loan checking and validation should undergo rigorous validation process. Without any doubt, this validation process should be done in an acceptable time limit with no delay affecting the bank staff and applicants [2].

In the conventional loan approval process, the bank hires an expert to approve/disapprove the loan applications by adopting the 5C principle (i.e., character, capacity, collateral, Capital, and Condition). The conventional methods may suffer of subjectivity and inaccuracy as the loan assessor will use the personal experience and knowledge to judge [5]. Another approach in loan approval is to decide the loan applicant eligibility through credit risk assessment process whereby on a numerical score called 'Credit Score' will be computed based on credit history , payment history, and background [6]. Kulkarni and Dhage [7] defined the credit score as a numerical value with three digits projected to grasp the eligibility of loan applicant based on his/her financial history. However, credit scoring

Manuscript received January 5, 2025
Manuscript revised January 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.1.21

requires both experts alongside statistical algorithms to decide the applicant eligibility.

Furthermore, as the loan approval process needs diverse information from different entities (e.g., government, other banks), the need for accurate, efficient, and autonomous loan approval model is very high. Therefore, quite recently, the researchers and the banking authorities have opted for training classifiers based on various machine learning and deep learning algorithms to automatically predict the credit score of an applicant based on their credit history and other historical data and make the process of selecting the eligible candidates a lot easier before the loan is approved [5-6]. For example, to predict the eligibility of loan applicant, machine learning algorithms are widely used. Therefore, the main aim of this is paper is to predict the loan default based on different loan factors or features. Particularly, a public dataset was used in this research to predict whether the loan applicant will be defaulter or not. This loan prediction problem is a binary classification to get the Loan_default Status value as yes or no. To achieve the paper aim, four machine algorithms were trained and tested (i.e., K-Nearest Neighbor, Logistic Regression, Decision Tree, and Random Forest) on the dataset. The experiment starts with exploratory data analysis to get better understanding of the features, and their correlation. Then, the preprocessing techniques are used to treat the missing values and outlier.

The rest of the paper is structured as follows. Section two discusses the background and related works, section three demonstrates the dataset used and the implementation details, and section four records and discusses the results.

2. Background and Related Work

Approving the financial loan properly either online or physically is a shared responsibility to protect the bank assets, the client rights, and the national economy. Such a responsibility prevents any financial loss due to the loan default. Thus, there is emphasis on the importance of minimizing the loan default. The banks and debaters used to rely on financial experts to judge the loan eligibility based on their experience. In addition, credit scoring was introduced in the loan processing to assess the applicant's ability to pay the incurred financial obligations. In credit scoring, the assessors predict if the loan applicant is obligated to pay his dues based on many factors including the available credit, the loan amount, the loan rate, and the applicant profile [8].

Furthermore, to assess the credit score of loan applicant accurately, diverse information from different parties, including the competent banks and governmental entities, are required to build a complete user profile. This is why the banks tends to collaborate through sharing the needed information that help in making a loan approval decision. However, the data needed for credit scoring are huge and may contains unnecessary data, which affects the assessment process negatively [9]. Therefore, using data analysis techniques in banking helps a lot in to make an accurate decision in the loan approval process.

In the literature, there are many works focusing on using data analysis, particularly machine learning, to predict the loan status. However, there are many other techniques also used in predicting the loan defaulters such as Artificial Neural Networks (ANN) [10-11], fuzzy logic [12], Genetic Algorithms [13-14]. Bae and Kim [15] and Arun et al. [16] proved the importance of Machine learning to validate the eligibility of loan applicant in order to minimize the credit risk factors. As an example of using machine learning , Vaidya [19] uses logistic regression model to predict the loan approval based on a set of records of loan applicant. Although Logistic regression is able to run with the nonlinear data, it requires independent variables for estimation and a large sample is required for parameter estimation [17]. Moreover, Turkson et al. [18] evaluated around 15 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 3 selected features. Similar work found in [19] whereby the logistic regression was used and tested on independent (non-correlated) features. Similarly, Kemalbay and Korkmazoğlu [20] used a logistic regression model with uncorrelated components to predict the housing loan approval of a private bank in Turkey.

On the other hand, decision trees were used by [13, 21, 22]. Somayyeh and Abdolkarim [13] used a decision tree model to calculate the rating for bank customers to reduce the credit risks in Mellat Bank of Iran, while Amin and Sibaroni [22] used Decision tree algorithm called C4.5 to implement a predictive a model. According to [22], this C4.5 algorithm gave a high precision and accuracy of 96.4% to predict loan defaulters on a dataset 1000 cases. In addition, Gradient Boost Ensemble algorithm was used by Lawi et al. [23] to predict the loan defaulter on German language- based dataset. The results showed a better accuracy with 81%. Similar work is proposed by Yadav et al. [21] whereby the decision tree and k-fold were used to split the loan applicants to different groups based on the most differentiator variables. Another work was presented by Priya et al. [24] that used the random forest to build loan prediction model whereby the features were analyzed to prove that credit history is the most important feature that influences the loan approval. The achieved accuracy by Priya et al. [24] was 81%. Another work by Odegua [25] used gradient boosting algorithm to build a loan prediction model. The proposed model achieved an accuracy of 0.79 with two discriminating features which are geographic location and applicant age.

On the other hand, Taneja et al. [26] used fuzzy rules to analyze the loan factors such as job status, weight, and income source on realistic dataset obtained from the bank of England. Taneja et al. [26] compared their works with different works and achieved accuracy of 83% on the given dataset. Another direction was followed by Jiang et al. [27] whereby Jiang et al. analyzed the descriptive text of the loan to extract some of the soft features (e.g., Part-of-Speech features, sentiment features, and social relationship information) that can be integrated with the traditional features to improve the prediction accuracy. The soft features were extracted using Latent Dirichlet Allocation (LDA).

Another direction of using machine learning is to compare multiple Machin Learning algorithms to find the best algorithm based on the used data. For instance, Gahlaut and Singh [28] proposed prediction model to decide the loan approval based on some factors such as occupation, financial position, and family background. Gahlaut and Singh used two algorithms named Random Forest (RF) and linear regression (LR). The results showed that RF outperforms LR with accuracy values 0.79 and 0.76 respectively. Likewise, Arora and Kaur [29] tested four algorithms namely Random Forest, SVM, Naive Bayes and K Nearest Neighbors (KNN) to show how accurately they can predict the loan defaulters. They found that Random Forest (RF) algorithm gave better accuracy than the others if it is used with optimized feature selection methods. Another work was introduced by Zhou et al. [30] whereby five Machine algorithms, namely Random Forest, Decision Tree, Bayes classification, Bagging, and Boosting, were tested in the field of loan defaulter prediction. The results show that decision trees show better performance. In addition, the efficiency of decision tree can be improved using Boosting. Although Zhou et al. [30] introduced a good explanation for the used algorithms, the work lacks the experimental results, in-depth analysis, or model evaluation. Alternatively, Arya [8] conducted a comparison using data-driven approach for three algorithms named Decision Tree (DT), Artificial Neural Network (ANN), and support vector machine (SVM). The results showed that SVM provided higher accuracy (72.05%) than the others. Additionally, the results were validated using 10-fold cross validation.

Likely, Hamid and Ahmed [31] compared j48, bayesNet, and naive Bayes algorithms. The highest accuracy that they achieve is 78%. Moreover, Turkson et al. [18] evaluated around 15 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 3 selected features. To evaluate the used classification models, Song and Peng [32] proposed a detailed evaluation framework for the loan classification models used by banks.

Likewise, Metawa et al. [14] conducted a comparative study using different machine learning algorithms to predict the mortgage loan 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. Besides, Kemalbay and Korkmazoğlu [20] applied LightGBM, XGBoost, Random Forest, and Logistic Regression algorithms to build a binary prediction model that predict the loan defaulters. Random Forest showed better performance with AUC value of 0.89 and accuracy value of 0.88. In addition, the proposed evaluation model is based on multiple criteria decisions making (MCDM) whereby the performance of the different model will be evaluated against a set of a set of performance metrics to select the high rank prediction model.

Based on the above discussion, we can conclude that there are diverse Machine Learning algorithms used in Loan Defaulter prediction, however, most of the studies considered the accuracy, recall, precision values. Further information on the three algorithms can be found in previous study [35]. However, further investigation still needed to validate the performance of these algorithms. Therefore, this paper aims to conduct a comparative analysis to investigate the classifiers' behavior for using additional metrics: AUC and ROC. Receiver Operating Characteristics (ROC) curve, which is a graph that shows the performance of the classification algorithm at all classification thresholds [36]. The ROC evaluation metric measures the performance of the model in different points while the model is operational, and it gives a graphical representation of the classification model performance. The ROC curve is a graphical representation of TPR versus FPR [37-38]. The ROC curve is an important method to compare classifiers visually, but it doesn't give us a metric value by itself, and the Area Under Curve (AUC) gives that as it is covering the area under the blue scribbly. [36, 39]. ROC and AUC provide different performance indications of different classifiers to determine that best one. Each one of them provides different representations such as a matrix, graph, or number.

In addition, throughout the above discussion, we notice that most of the works in literature showed a good result of Logistic regression, Decision Tree, Random Forest, and to some extent K-Nearest Neighbor (KNN) algorithms. Thus, in the following sub-section, we will introduce those ML algorithms as a focus of this study.

2.1. Machine Learning Algorithms

K-Nearest Neighbour (KNN): KNN is used to solve classification problems based on stored data. The algorithm trains the dataset and stores it in the memory. When the classification process is to test new data points, the algorithm works on the basis of similarity of the state

between the new data point and the stored dataset and classifies new data in accordance with the most similar class on the basis of the value of K and the closest one on the basis of Euclidean distance.

Logistic regression is a statistical technique that is used to predict the probability of binary response variables. It is used when our label(y) is a binary response variable in the form of 1 or 0, yes or no, etc. It is a basic and popular algorithm to solve a classification problem.

Decision Tree classifier. Decision tree is a popular algorithm that is used to build classification models. The models are built in the form of a tree-like structure. Each node in the tree indicates a test on a variable, and each branch descending from that node indicates one of the possible values for that attribute. The third algorithm is Random Forest Classifier.

Random Forest is a classification algorithm that consists of many decision trees. It tries to create an uncorrelated forest of trees by using feature randomness and Bagging. The prediction of an uncorrelated forest of trees is more accurate than that of any individual tree.

3. The proposed Loan Default Prediction Methodology

In this research, the main goal is to predict if the loan applicant will be a loan defaulter based on the historical data provided. The loan will be labeled as "default loan" or "non-default loan" based on the classification model that we use.

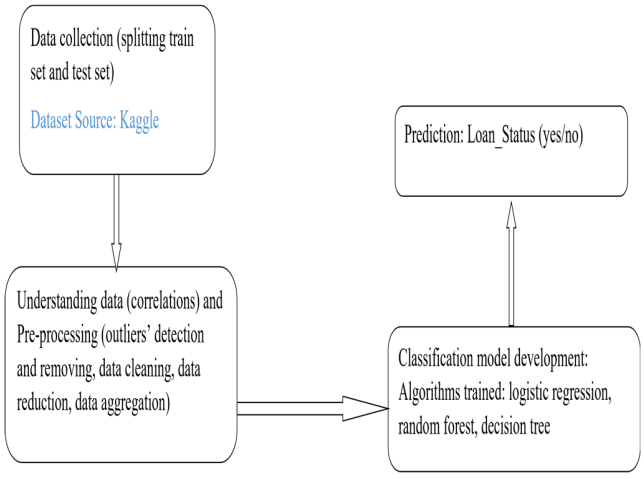

Fig. 1: The proposed Loan prediction methodology

Thus, we adopted the methodology shown in Fig.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]. The dataset used shows two sets: accepted loans and rejected loans for the period between 2007 to 2020 Quarter1.

Then, the data exploration phase was conducted whereby we tried to understand the data and relationships between features through different types of tables and figures such as box plots, bar plots, histograms, heatmap, etc. The third phase is to train three classification algorithms on the explored data based on the best-found variables that may be helpful in the prediction process.

3.1 Dataset Description and Preprocessing

In order to validate the proposed model and benchmark the results with other models, we have used a public dataset for loan default prediction by Lending club, as explained earlier. The used dataset consists of more than 30 features. In this paper, we dropped some features that are irrelevant to the loan prediction or have no impact on the target variable such as issued, grade, sub_grade, emp_title, and addr_state. The main feature that we adopt in this paper is described, with their accompanying descriptions, in Table 1. The dataset was partitioned to training set (80%) and the reaming (20%) for the testing purpose. The dataset, then loaded into the Google Colab notebook by mounting the drive.

3.2 Data pre-processing

Missing values are a serious issue in any dataset. Thus, in the data pre-processing stage, we started checking if there are any missing values may affect the prediction, so the heatmap was used. As depicted in Fig. 2, heatmap shows that there are missing values in different features. From the heat map, we can guess initially that the majority of missing values exist in emp_title, emp_length, revol_util, mort_acc features.

Hence, we need to analyze those features in detail to know how to deal with the missing values. As depicted in Fig. 3, the missing values in the features: emp_length, and mort_acc where the emp_title feature and title feature were dropped. To deal with the missing values, statistical measures, such as mean, median and standard deviation, were applied to impute the missing values. In this case, the mean method was applied to replace the numerical values

for both emp_length, and mort_acc by calculating the mean for the features and replacing the missing value.

Table 1: Dataset introduction

Variable	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
loan_status	Current status of the loan
revol_bal	Total credit revolving balance
annual_inc	The self-reported annual income provided by the borrower during registration.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
mort_acc	Number of mortgage accounts.
pub_rec_bankruptcies	Number of public record bankruptcies
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

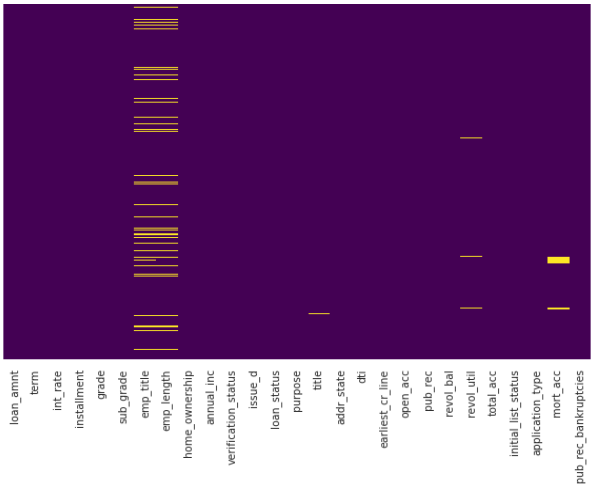

Fig. 2: Heatmap technique for missing feature values discovery

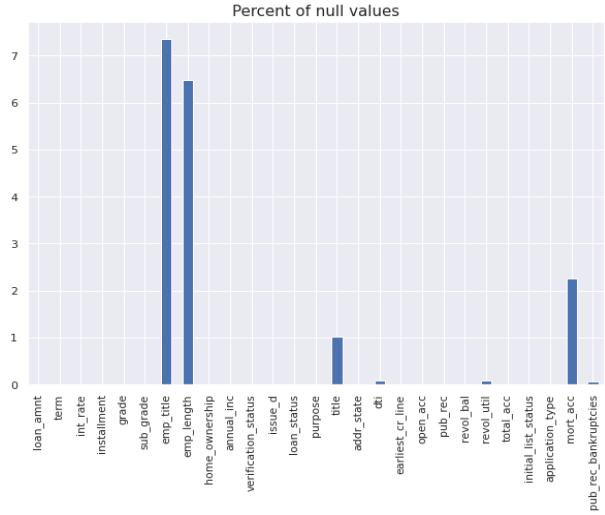

Fig. 3: The percent of null values in the different features.

After treating the missing values for the concerned features, we have to check if there is any outlier in the processed data. Box- The next step after checking the missing values is to find if there are any outliers in the collected data. box plots technique was used for this purpose, as depicted in Fig. 4. The box plots in Fig. 4 shows that there are four features with outliers, which are annual_inc, loan_amnt, term, and installment. After detecting the outliers, the univariate method was applied to remove the outliers.

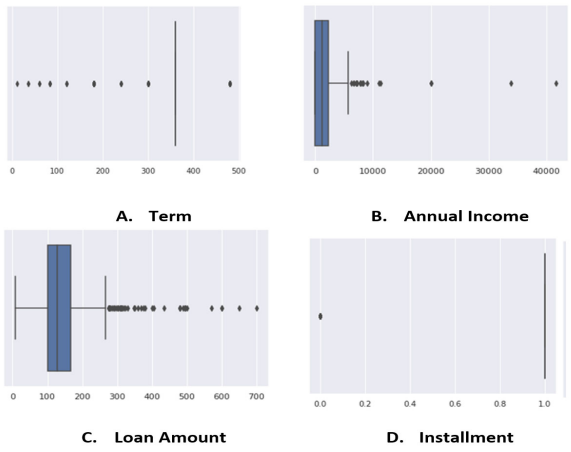

Fig. 4: The Outlier Detection Using Box Plots Techniques for Different Features

The last step in the preprocessing is to find the most influential features in the prediction process. This can be done by finding the correlation between the data features using heat map to visualize the features correlation. Fig. 5 depicts the heat map for the collected data attributes whereby we can notice easily that the most important feature for loan prediction.

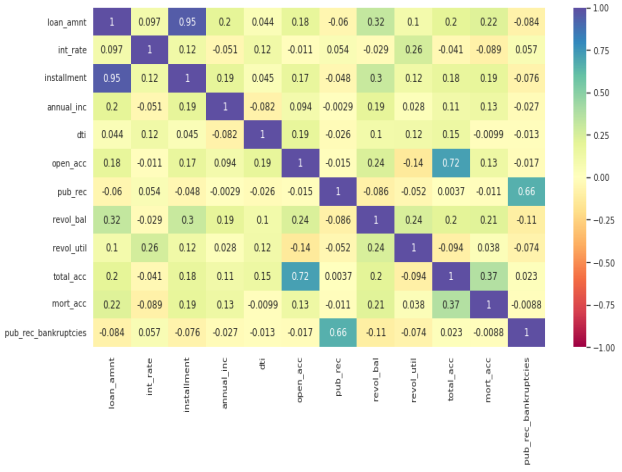

Fig. 5: Heat map Presentation to show the features correlation.

Another step to be done here is to check the data imbalance to ensure that the prediction model will give accurate results. At this stage, the dataset was checked to show the class imbalance. It is obvious in Fig. 6 that the given dataset is imbalanced where the default class is a minority with very less observations. Such a minority may affect the prediction stage to make the classifier biased in favor of the non-default class. Therefore, the class imbalance issue should be considering during the prediction stage.

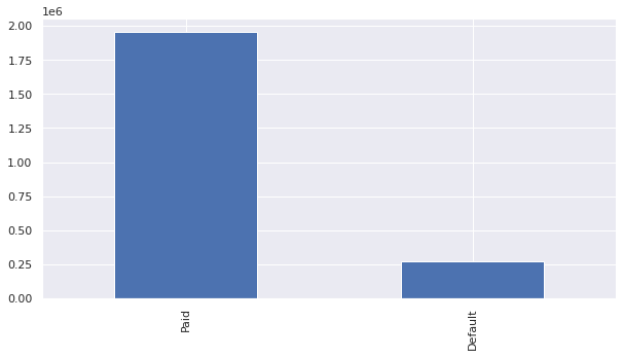

Fig. 6: The class imbalance in Loan Default Dataset

3.3 The proposed prediction Module

After performing pre-processing, various classifiers have been used to predict whether the loan is default or not. The classifier models constructed are K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR), as discussed in section 2.

As we mentioned earlier, the dataset suffers of class imbalance, thus we need to treat this issue during the experiment to show if this issue affects the results or not. Actually, we use a combination of under sampling and oversampling techniques. We will elaborate those techniques in the following lines.

3.4. Data imbalance & Resampling Techniques

In order to address the class imbalance, the random resampling approach is used to rebalance the class distribution. There are two mechanisms namely, under-sampling and over-sampling. 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. The advantage of under-sampling is reducing the cost of learning process; however, it has some limitations like increasing the class variance that give wrong prediction [41] or ignoring important examples that are necessarily needed by the algorithm to learn [42]. Some of the under-sampling techniques are Random under-sampling (RUS), Condensed Nearest Neighbor (CNN), Tomek, One Sided Selection (OSS), Edited

Nearest Neighbors (ENN), Repeated Edited Nearest Neighbors (RENN), All K-Nearest Neighbors (AllKNN), Neighborhood Cleaning Rule (NCR), Near Miss (NM), and Instance Hardness Threshold (IHT). Therefore, oversampling techniques aim to replicate the minority class instances to rebalance the dataset.

However, such techniques may lead to poor performance of the model in some cases because it may be hard to generate the minority data in the training set [43-44]. Examples of oversampling techniques are SMOTE, Random Over Sampler, Adaptive Synthetic (ADASYN), Borderline SMOTE (Border), and SVM SMOTE (SVM). Another approach to taggle the class imbalance is to combine both under-sampling and over-sampling techniques together in the training set such as SMOTEENN (SMENN) and SMOTETomek (SMTOMEK). In this paper, we aim to minimize the inconsistency between the percentage of non-target and target cases in the majority class using under-sampling to make the ratio of non-target to target cases 80:20. Then we will apply the over-sampling on the minority class using replacement to replace 50% of the default cases.

4. Results and discussion

This section records the experimental results of running the proposed prediction model using the four machine learning algorithms (KNN, DT, RF, and LR), and then record the results of combining those algorithms with resampling techniques (under-sampling (US), and over-sampling (OS) . The performance metrics used are precision, Recall, F1, Accuracy, and Area under curve (AUC). Table 2 summarizes the Performance Summary of the classifiers. LR outperformed the others in terms of accuracy (81%), F1 score (88%), and AUC (65%). However, KNN showed better precision and recall values.

Another experiment was conducted to show the performance of the four algorithms with resampling. We apply under-sampling, then we apply the combination of under-sampling and over-sampling. Table 3 summarizes the results of applying the resampling on the four algorithms. As shown in the table, applying the under sampling has a very minor effect on the results with slight improvement. However, the results have been improved significantly when the combination of both under-sampling and over-sampling techniques applied. The experimental results showed the superiority of LR algorithm , even with applying the resampling on the other algorithms.

Table 1. Summary Results of the Four machine learning algorithms

Algorithms	Precision	Recall	F1	Accuracy	AUC	ROC
KNN	0.71	0.84	0.75	0.78	0.69	0.61
LR	**0.70**	**0.70**	**0.88**	**0.81**	**0.70**	**0.68**
DT	0.72	0.75	0.80	0.74	0.65	0.67
RF	0.74	0.93	0.84	0.79	0.69	0.69
Results for Random under-sampling						
Algorithms	Precision	Recall	F1	Accuracy	AUC	ROC
KNN	0.71	0.84	0.75	0.78	0.69	0.62
LR	0.74	0.78	0.89	0.84	0.72	0.69
DT	0.72	0.73	0.80	0.74	0.65	0.67
RF	0.74	0.88	0.84	0.79	0.71	0.70
Results for combing under-sampling & oversampling						
Algorithms	Precision	Recall	F1	Accuracy	AUC	ROC
KNN	0.83	0.90	0.87	0.94	0.76	0.67
LR	**0.89**	**0.98**	**0.90**	**0.93**	**0.80**	**0.72**
DT	0.81	0.85	0.89	0.89	0.75	0.71
RF	0.84	0.92	0.87	0.90	0.79	0.74

In order to compare our work with the others, we choose the model proposed by Coser et al. [34]. Coser et al. [34] compared different algorithms with resampling on the same dataset. The comparison results are recorded in Table 4. The performance metric used in [34] is AUC, thus, we will show the comparison results in terms of AUC as per the information available in [34]. We can notice that LR and DT gave better results in work better than the results in [34], however, the two LightGBM and XGBoost are still better.

Table 2. Result comparison

	The algorithm	AUC (without resampling)	AUC (with resampling)
Coser et al. [34]	LightGBM	0.75	0.89
	XGBoost	0.73	0.75
	RF	**0.63**	**0.74**
	LR	**0.57**	**0.60**
Our model	KNN	0.69	0.76
	LR	**0.70**	**0.80**
	DT	0.65	0.75
	RF	**0.69**	**0.79**

Conclusion

One of the critical businesses for any financial organization is the loan processing. Thus, the banks aim to approve the loan in a proper way to avoid any negative impact. Loan default is one of the critical financial issues whereby the borrower will not be able to pay back the financial obligations. To avoid such issues, the bank tries to minimize the errors in loan approval process to ensure approving the loan for the deserving applicants through hiring loan approval experts. However, the data needed for this process is huge and needs a careful attention from the bank. Thus, machine learning was introduced for this context. In this paper, we proposed a prediction model to predict the loan defaulter based on different relevant features. The proposed methodology consists of preprocessing to clean the data, removing the outliers, and then finding the correlation between the features to find the most influential feature. Then, four machine learning algorithms were deployed with resampling techniques to be trained and tested. The novelty of this research can be represented in comparing four machine learning algorithms with different resampling techniques to find the most accurate prediction. The experimental results showed the superiority of Logistic Regression in combination with resampling techniques on the other three algorithms in terms of accuracy precision, Recall, F1, and Area under curve (AUC). In the future direction, we will work on comparing more machine learning algorithms to come out with better accuracy. Besides, we can improv ethe model accuracy throughout applying the ensemble machine learning algorithms and also applying feature extraction.

References

[1] Fati S. M. Machine Learning-Based Prediction Model for Loan Status Approval. Journal of Hunan University Natural Sciences. 2021, 48(10).

[2] Gupta, A., Pant, V., Kumar, S. and Bansal, P.K. Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) 2020, 423-426.

[3] Kohv, K. and Lukason, O. What Best Predicts Corporate Bank Loan Defaults? An Analysis of Three Different Variable Domains. Risks, 2021, 9(2), 29.

[4] Aphale, A.S. and Shinde, S.R. Predict loan approval in banking system machine learning approach for cooperative banks loan approval, International Journal of Engineering Research & Technology, 2020, 9(8), 991-995.

[5] Madaan, M., Kumar, A., Keshri, C., Jain, R. and Nagrath, P. Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering, , 2021, 1022(1), 012042.

[6] Ahmed, M.I. and Rajaleximi, P.R. An empirical study on credit scoring and credit scorecard for financial institutions. Int. Journal of Advanced Research in Computer Engineering & Technol.(IJARCET), 2019, 8, 275-9.

[7] Kulkarni, S.V. and Dhage, S.N. Advanced credit score calculation using social media and machine learning. Journal of Intelligent & Fuzzy Systems, , 2019, 36(3), .2373-2380.

[8] Arya, S., Eckel, C. and Wichman, C. Anatomy of the credit score. Journal of Economic Behavior & Organization, 2013, 95, 175-185.

[9] Tripathi, D., Edla, D.R., Bablani, A., Shukla, A.K. and Reddy, B.R. Experimental analysis of machine learning methods for credit score classification, Progress in Artificial Intelligence, 2021, 1-27.

[10] Yang, B., Li, L.X., Ji, H. and Xu, J. An early warning system for loan risk assessment using artificial neural networks. Knowledge-Based Systems, 2001, 14(5-6), 303-306.

[11] Hassan, A.K.I. and Abraham, A. Modeling consumer loan default prediction using ensemble neural networks. In 2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), 2013, 719-724.

[12] Kumar, B., Bawane, I., Shirsathe, A. and Pardeshi, P. An Expert System Based on Fuzzy Logic for Automated Decision Making For Loan Approval. International Journal of Recent Advances in Multidisciplinary Research, 2016, 02(12), 1078-1082.

[13] Somayyeh, Z. and Abdolkarim, M. 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, 2015, 3(2), 307-316.

[14] Metawa, N., Hassan, M.K. and Elhoseny, M. Genetic algorithm based model for optimizing bank lending decisions. Expert Systems with Applications, 2017, 80, 75-82.

[15] Bae, J.K. and Kim, J. A personal credit rating prediction model using data mining in smart ubiquitous environments. International Journal of Distributed Sensor Networks, 2015, 11(9), 179060.

[16] Arun, K., Ishan, G. and Sanmeet, K. Loan approval prediction based on machine learning approach. IOSR Journal of Computer Engineering, 2016, 18(3), 18–21.

[17] Khan, A., Bhadola, E., Kumar, A. And Singh, N. Loan Approval Prediction Model A Comparative Analysis. Advances and Applications in Mathematical Sciences, 2021, 20(3), 427-435.

[18] Turkson, R.E., Baagyere, E.Y. and Wenya, G.E. A machine learning approach for predicting bank credit worthiness. In 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), 2016, 1-7.

[19] Vaidya, A. Predictive and probabilistic approach using logistic regression: application to prediction of loan approval. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, 1-6.

[20] Kemalbay, G. and Korkmazoğlu, Ö.B. Categorical principal component logistic regression: a case study for housing loan approval. Procedia-Social and Behavioral Sciences, 2014, 109, 730-736.

[21] Yadav, O., Soni, C., Kandakatla, S. and Sawant, S., Loan Prediction System Using Decision Tree. International Journal of Information and Computing Science, 2019, 6(5), 137-143.

[22] Amin, R.K. and Sibaroni, Y. 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), 2015, 75-80.
[23] Lawi, A., Aziz, F. and Syarif, S. Ensemble GradientBoost for increasing classification accuracy of credit scoring. In 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 2017, 1-4.
[24] Priya, K.U., Pushpa, S., Kalaivani, K. and Sartiha, A. Exploratory analysis on prediction of loan privilege for customers using random forest. Int. J. Eng. Technol, 2018, 7(2.21), 339-341.
[25] Odegua R. Predicting bank loan default with extreme gradient boosting. arXiv Prepr. arXiv, 2002,02011, 2020.
[26] Taneja, S., Suri, B., Gupta, S., Narwal, H., Jain, A. and Kathuria, A. A fuzzy logic based approach for data classification. In Data engineering and intelligent computing, 2018, 605-616.
[27] Jiang, C., Wang, Z., Wang, R. and Ding, Y. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research, 2018, 266(1), 511-529.
[28] Gahlaut, A. and Singh, P.K. Prediction analysis of risky credit using Data mining classification models. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, 1-7.
[29] Arora, N. and Kaur, P.D. A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 2020, 86, 105936.
[30] Zhou, P.Y., Chan, K.C. and Ou, C.X. Corporate communication network and stock price movements: insights from data mining. IEEE Transactions on Computational Social Systems, 2018, 5(2), 391-402.
[31] Hamid, A.J. and Ahmed, T.M. Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal (MLAIJ), 2016, 3(1), 1-9.
[32] Song, Y. and Peng, Y. A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. IEEE Access, 2019, 7, 84897-84906.
[33] Bagherpour, A. Predicting mortgage loan default with machine learning methods. Ph.D. dissertation, University of California, Riverside, 2017.
[34] COŞER, A., Maer-matei, M.M. and ALBU, C. Predictive Models for Loan Default Risk Assessment. Economic Computation & Economic Cybernetics Studies & Research, 2019, 53(2), 149-165.
[35] Muneer A. and Fati S. M. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet, 2020, 12(11), 187-200.
[36] "Classification: ROC Curve and AUC" | Google Developers, 2021. [online]. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, Accessed on 22- Nov- 2021.
[37] Padhi, B.K., Chakravarty, S. and Biswal, B.N. Anonymized credit card transaction using machine learning techniques. In Advances in Intelligent Computing and Communication, 2020, 413-423.
[38] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S. and Kuruwitaarachchi, N. Real-time credit card fraud detection using machine learning. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, 488-493.
[39] Sarkar D., Bali R.and Sharma T. Practical Machine Learning with Python. A Problem-Solvers Guide to Building Real-World Intelligent Systems, Berkely: Apress, 2018.
[40] "Deep Learning & Analysis: Loan Default Prediction" | Kaggle 2021. https://www.kaggle.com/slythe/deep-learning-analysis-loan-default-prediction. Accessed on 01-Aug – 2021.
[41] Dal Pozzolo, A., Caelen, O. and Bontempi, G. When is undersampling effective in unbalanced classification tasks?. In Joint european conference on machine learning and knowledge discovery in databases, 2015, 200-215.
[42] Wasikowski, M. and Chen, X.W. Combating the small sample class imbalance problem using feature selection. IEEE Transactions on knowledge and data engineering, 2009, 22(10), 1388-1400.
[43] García, V., Mollineda, R.A. and Sánchez, J.S. On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications, 2008, 11(3), 269-280.
[44] Cieslak, D.A. and Chawla, N.V. Start globally, optimize locally, predict globally: Improving performance on imbalanced data. In 2008 Eighth IEEE International Conference on Data Mining, 2008, 143-152.

Suliman Mohamed Fati. Assistant Professor in Information Systems Department, CCIS, Prince Sultan university, Saudi Arabia. He obtained his Ph.D. (2014) from Universiti Sains Malaysia (USM) – Malaysia. His research interests focus on Internet of Things, Machine Learning, Cloud Computing, Cloud Computing Security, and Cybersecurity. He is a IEEE senior member, and also an active member in different professional bodies as ISACA, IACSIT, IAENG, and Institute of Research Engineers and Doctors, USA. He serves as reviewer in many international impact-factor journals, and a TPC in different international conferences.