Jurnal Kejuruteraan 36(5) 2024: 1935–1944
https://doi.org/10.17576/jkukm-2024-36(5)-13

Algorithm Comparison for Data Mining Classification: Assessing Bank Customer Credit Scoring Default Risk

Elaf Adel Abbas[a*] & Nisreen Abbas Hussein[b]

[a]*College of Computer Science and Information Technology,University of Karbala*

[b]*Babylon Education Directorate, Ministry of Education, Iraq,*

**Corresponding author: elaf1982adil@gmail.com*

Received 22 January 2024, Received in revised form 13 May 2024
Accepted 11 July 2024, Available online 30 September 2024

ABSTRACT

Rating consumer credit risk involves assessing credit application risks. Thus, every business must appropriately identify debtors and non-debtors. This study uses machine learning approaches to simulate consumer credit risk and compares the results to the logistic model, determining if machine learning improves client default ratings. The study examines how customer attributes affect virtual experiences. Despite advances in machine learning models for credit assessment, unbalanced datasets and some algorithms' failure to explain forecasts remain major issues. This study used 2005 Taiwanese credit card consumers' education, age, marital status, payment history, and sex. The default experience is modeled using Logistic Regression, K neighbors, Support Vector Machine, Decision Tree, Random Forest, Ada Boost Classifier, and Gradient Boosting. The models' Accuracy, precision, recall, receiver operating characteristic (ROC) curve, and precision-recall curve were evaluated. Random Forest's 97% ROC metric rating outperformed all other accuracy metrics. The logistic model underperformed, while machine learning improved the default categorization.

Keywords: Credit scoring; artificial intelligence; machine learning; classification techniques; logistic regression

INTRODUCTION

Financial risk management is a delicate topic that should be investigated. Some organizations, industries, and governments worldwide depend on risk management systems and credit scoring (Zhou et al. 2018). Financial fraud, which includes business fraud, personal loan fraud, money laundering, credit card fraud, insurance fraud, peer-to-peer lending fraud, and others, is a conscious strategy, culpability, or fraud committed with the intention of exploiting the structure of a nonprofit organization in order to illicitly achieve financial benefit without resorting to physical coercion (Pławiak et al. 2020). The line separating fraudulent activity from damaging credit events is becoming hazier in the credit markets as more credit events shift online, and counterfeiters improve their skills. As a result, financial institutions frequently combine financial detection, credit scoring, and other factors when making decisions in order to lower the risk of credit loss. Banks and customers face many risks. Banks use Credit Scoring (CS) to evaluate loan applications(Durand, 1941). To manage financial risks and decide whether to lend money, banks and other financial institutions must collect customer data. This method can help identify good and bad debtors. Banks consider "good borrowers" clients with clean credit histories. "Bad borrowers" have poor credit. A simple selection technique may not always classify correctly. More accurate automated approaches that reduce prediction errors are critically needed to manage vast and complicated CS datasets (Anderson 2007). Model development and model implementation comprises the credit rating process. The first step is to collect samples of good and bad loan applications from past borrowers to train and construct a model that can predict payment behavior. Formally, let $A = \{a_i, b_i\}$, where a_i is the number of loan applications and b_i represents their status as good or bad loans. The loan application form has several properties or variables $a_i = (a_{i1}, a_{i2},..., a_{im})$. Thus, a quantitative model is constructed to convert loan application characteristics to the chance of default[5]. After the model's development and training, it's time to test it and see how well it classifies loan applicants.

The applicant's final score, which the lender will use to decide whether to grant the loan, is based on a threshold or drop score of Threshold value (Tc). A loan applicant's status is usually (0) for good and (1) for bad. The model's score is f(x) for new loan applications. If this score is below Tc, the loan is approved; otherwise, it is denied.

LITERATURE REVIEW

This section specifically highlights how financial institutions are developing and implementing cutting-edge technologies based on Artificial intelligence- Machine Learning (AI-ML) strategies to deal with their various credit risks in both developed and emerging nations. The majority of financial organizations today deal with various risks daily. Credit risk, operational risk, market risk, and liquidity risk are a few of these dangers (Leo et al. 2019).

Few writers have discussed the socioeconomic implications for determining a client's credit score in earlier research papers, which have mostly focused on a customer's demographics and statistical factors (Moradi & Mokhatab Rafiei, 2019). The authors emphasized that political alterations have an impact on economic aspects as well. In order to estimate credit risk, they also took politico-economic issues into account. To first anticipate whether a specific loan is performing, they created an adaptive network-based fuzzy inference system. Most banks and other financial institutions now prioritize social and economic effects due to Covid-19.One of the writers evaluated their customers' credit scores using data from an Iranian bank, especially when societal and economic conditions are exceptional. Using the features of Iranian bank clients' behavior as input, that assessed credit scoring using a fuzzy inference method, outperforming more traditional models, especially during economic crises.

Researchers have stressed the non-linear and non-parametric correlations between the factors influencing bank lending and how many loans are still outstanding (Ozgur et al. 2021). They showed how 19 macroeconomic, local, and international variables impacted Turkish bank loans between 2002Q4 and 2019Q2. They contrasted the regression model with ML-based approaches to determine how these factors will affect the results. The authors also pointed out that conventional linear regression methods struggled with the extremely high dimensionality of the datasets, whereas ML-based techniques were able to accommodate this. For the majority of their debt recovery management, banking institutions rely on outside sources, which entails increased expenses and market risks. Therefore, it is always advised to have a reliable strategy in place for predicting debt repayment before extending any credit to the debtors.

Some authors classified the sufficiency of the borrower using ML approaches like Random Forest (RF) and AdaBoost (Aniceto et al. 2020). Using the loan database from the Brazilian Bank, researchers examined various ML techniques and evaluated the suitability of the borrowers. Large Brazilian financial institutions' low-income borrowers make up the majority of the data sets. The portfolio's default rate was close to 48%. Using real-world data, they developed a machine learning (ML) model and showed that Random Forest and AdaBoost performed better than competing methods. Only some authors suggested using the model of decision trees to determine whether the loan provider poses a risk for performing or non-performing loans. Most academics emphasized that a categorization issue exists with credit scores (Boughaci & Alkhawaldeh, 2018). They compared the credit data sets from Germany and Australia with well-known classifier benchmarks. They coupled the Support Vector Machine (SVM) model with the Local Search method (LS), Stochastic Local Search technique (SLS), and Variable Neighborhood Search (VNS) approach to determine a person's credit score.

MACHINE-LEARNING CLASSIFICATION TECHNIQUES IN CREDIT-SCORING

The major goal is to create a model that can effectively categorize and measure borrower repayment behavior as well as anticipate borrower loan applications. This section provides an overview of the most popular modern machine-learning categorization approaches that are pertinent to this research and utilized to create credit-scoring models.

LOGISTIC REGRESSION

A specific type of Generalized Linear Model (GLM), or "Logistic Regression (LR)," is a generalization of the ideas of regular linear models. As a result, logistic regression is similar to linear regression and is employed in this analysis to solve a classification problem. A binary outcome variable, typically denoted by 0 or 1, is modeled using LR.

According to Thomas, the scoring model's result must be binary (accept/good loan, 0; reject/bad loan, 1), and this is based on a number of independent variables(Ala'raj & Abbod 2015).

K NEAREST NEIGHBOR

K nearest neighbor is one of the most widely applied credit scoring techniques (KNN). The non-parametric classification method category includes this technique. It is well known that the non-parametric classifier frequently experiences outliers, especially when the training sample size is small. Numerous credit scoring researchers have utilized KNN to evaluate the risk involved in making a loan to a business or a person(Mukid et al. 2018). The Euclidean distance between the given training samples and the test sample is a common foundation for the k-nearest neighbor classifier. The primary principle of the k-NN method is that the training data is used to select the k nearest neighbors of each new point that needs to be predicted. The average of the values of the new point's k-closest neighbors can then be used as a prediction (Zhang & Wang 2016).

DECISION TREES

Decision trees are now frequently used to fit data, anticipate default, and improve credit rating. The algorithms used by decision trees work top-down, selecting the variable that divides the dataset "best" at each stage. Any of a number of criteria, such as the Gini index, information value, or entropy, can be used to determine what is "best." Predict an outcome by following the tree's branches from the starting (root) node to a leaf node. The final leaf node contains the solution. Classification trees deliver nominal responses, such as "true" or "false," rather than more nuanced answers. Regression trees produce numerical results (Bastos, 2007).

RANDOM FOREST CLASSIFICATION

Random forest is a machine learning technique for dealing with classification and regression problems. It employs ensemble learning, a technique for resolving challenging problems by integrating a variety of classifiers. It's a classifier that combines several decision trees on various dataset subsets and averages the outcomes to enhance the dataset's anticipated Accuracy. One potential benefit of these techniques is that they may allow model builders to significantly shorten the time spent on data management and data pre-processing (Tang et al. 2019).

SUPPORT VECTOR MACHINE

Another effective machine-learning method utilized in categorization and credit scoring issues is SVM. Due to its excellent outcomes, it is widely employed in the field of credit scoring as well as other areas. SVMs that take on the appearance of a linear classifier. SVMs predict a set of two classes of inputs to identify which of the two classes is most likely to have the output. In order to create the finest hyperplane (Line) that divides the input data into two groups, binary classification is accomplished using SVMs (good and bad credit).

SVM can be used in both linear and non-linear separation settings. The latter uses a basis expansion h(x) that can be converted back to a non-linear boundary in the original space to construct the linear boundary in an extended and revised version of the feature space. It is necessary to understand how the kernel function K computes the inner products of vectors in the transformed space by using the original space X as input (Dastile et al. 2020).

Fitting linear classifiers and suppressors with convex loss functions, such as those of (linear) Support Vector Machines and Logistic Regression, is a breeze with stochastic gradient descent (SGD). Text categorization and NLP are two areas where SGD has been successfully applied to tackle sparse and massive machine learning challenges.

The SGD does not belong to a particular family of machine learning models and is necessarily correct, just an optimization technique. The only purpose is to train a model. A straightforward stochastic gradient descent learning procedure that supports various classification penalties and loss functions is implemented by the class SGD Classifier (Condori-Alejo et al. 2021).

METHODOLOGY

The various steps of our process are depicted in Figure 1. Selecting data sets, cleaning and organizing information, developing models, and checking those models are all parts of the process described here. This study proposes methods to classify borrowers according to their payment risk in order to help banks avoid the real risks associated with non-payment, which are burdensome to banks. By focusing solely on trustworthy borrowers, banks may boost their profitability.

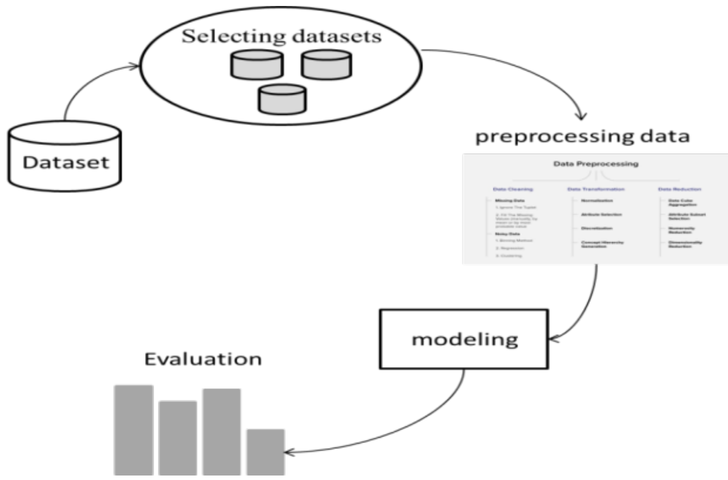

FIGURE 1. The procedure of credit cart

THE DATASET

The dataset chosen to be used for this study includes 30,000 Taiwanese credit card customers' anonymized information from 2005 variables as explanatory data (Yeh & Lien, 2009). The dataset includes characteristics of the clients, including whether or not they were in default on their obligations. Of the 30,000 debtors, 23 364 paid back their loans, while 6636 missed payments. Around 78 percent of the dataset's debtors are good debtors, and 22 percent are bad debtors. In this study, the result variable was a binary variable called default payment (Yes = 1, No = 0). As shown in Table 1, the data are divided into 23 columns with various numerical values and categorical information like education is also hidden as a numerical value.

TABLE 1. The Data Attributes

Attribute	Displaying
ID	ID of each client
Sex	Gender (1 = male; 2 = female)
Education	1: graduate school; 2 : university; 3 : high school; 4 : others
Marital status	1 : married; 2 : single; 3 : others
Age	Year
PAY1___PAY6	History of past payment. The past monthly payment records (from April to September, 2005) were tracked as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
BILL_AMT1__ BILL_AMT6	Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005
PAY AMT1__ PAY AMT6	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005.
DEFAULT	Default payment (1=yes, 0=no).

The magnitude of values varies between features due to some fields containing information about the account limit and payment details. All features were normalized to lessen the effects of this on the outcomes.

DATA PREPROCESSING

Data mining relies on several preparatory steps, one of which is called preprocessing. Some features may be redundant, while others are noisy and unimportant. Feature

Selection (FS) is a technique for figuring out the most helpful features. Preprocessing of the dataset is necessary when it contains useless data that is noisy (outliers), unreliable (missing), and inconsistent. Any extraneous and correlated data were eliminated from the dataset in order to increase model accuracy and obtain useful results. We next used data cleaning, discretization, and target class balancing to our data to produce a dataset that would work well with our algorithms.

DATA CLEANING

The methods employed to address missing data involve excluding or imputing the affected records with a predetermined value. In the case of noisy data, many techniques can be utilized, including binning algorithms, clustering, a combination of human and machine inspection, and regression analysis. Inconsistencies can be fixed manually. Some numbers in the datasets don't have official meanings on the UCI site; for example, the Education attribute's values ranged from one to four, but they also had values greater than four (331).

FEATURES SELECTION

A method to lessen dimensionality is feature selection. This method's primary use is the extraction of discrete subsets of pertinent features from the original dataset according to the assessment criterion.

DATA TRANSFORMATION

Data transformation is the process of converting data from one format to another so that it can be used for data mining. A few techniques for doing transformation include normalization, smoothing, aggregation, and generalization. The values in the dataset were all expressed as numbers in the records; for instance, Categorical information, such as sex, was encoded as a "1" for male and a "2" for female. That was problematic because including numbers in our data would make it less relevant to our clients; therefore, we needed to alter some columns to make them better suited for analyzing the outcome. We converted the string representations of these attributes (sex, education, and marital status).

DATA REDUCTION

Analyzing enormous amounts of data requires a lot of time. It can be accomplished using data cube aggregation, data compression, dimension reduction, data reduction, concept hierarchy, and discretization development. Because a group of academics came to the conclusion that discretization enhances the efficiency of the naive Bayesian algorithm, we discretized a continuous variable by applying it to our dataset (Jonathan L. Lustgarten, 2008.), Therefore, the Age property was split into ten-year chunks, and the Limit bal attribute was reclassified as Low, Medium, or High instead of ((0-100,000), (100,001-500,000), and over (500,001)) in New Taiwan dollars, to the labels.. Our next step in the preprocessing process was data reduction, where we shrunk the dataset to obtain two equal class representations, the default, and no default classes. From 30,000 records, we were able to reduce it to 13,210, which is a 50% reduction with 6,605 records for each class. To improve performance, we eliminated all redundant information from our dataset; as a result, there are now only six attributes instead of 24.

After that, a 70/30 split was performed randomly across the full data set to create a training set and a test set. This work uses an imbalanced dataset and employs sampling methods such as SMOTE, kNN, and Tomek-links. However, the study relied on a relatively simplistic random over-sampling method for the response variable due to technical limitations and time constraints. The model performance evaluation was conducted using the test data set, which was utilized to assess the models' predictive capabilities after training on the training set.

TECHNIQUES FOR CREDIT-SCORING EVALUATION

The following metrics are the most widely used metrics for evaluating the effectiveness of the models in credit scoring out of the many assessment metrics that are used in the many types of literature (Dastile et al. 2020). A confusion matrix is a common tool for assessing a classifier's performance (see Table 2). All of the instances in the data set are displayed in a confusion matrix and are divided into four groups:

TABLE 2. Confusion Matrixes Discretion

Observed class	Predicative Class	
	Class (1)=Good	Class (0)=Bad
Class (1)=Goods	True Positive	False Negative
Class(0)=Bad	False Positive	True Negative

TP (True Positive): These are the positive findings that the model correctly predicted based on the actual data, which, in our case, suggests that the model successfully predicted the number of defaults that actually occurred in the actual data.

TN(True Negatives): These are the negative values that the model correctly predicted based on the actual data, suggesting that the model correctly predicted the proportion of non-defaults that were non-defaults in the real data in our case.

FP (False Positives): These are the positive values the model mistakenly predicted based on the actual data, suggesting that the observations the model projected as default but were not in our case's real data.

FN (False Negative): These are the negative values that the model mistakenly predicted based on the actual data, suggesting that the observations the model projected as defaults weren't defaults in the real data in our case.

The consequences of misclassification in the context of credit rating are very different. False positives cause the lender to lose some or both the interest and the principal that was supposed to be repaid. In contrast, false negatives solely refer to the opportunity cost of lost interest that could have been gained. Due to the fact that these individuals were given a loan despite the model classifying them as excellent, false positives are substantially more expensive (Bunker et al. 2016) (Nyangena 2019).

The proportion of actual products that were accurately identified as such is known as the true positive rate:

$$TP = \frac{TP}{TP+FN} \tag{1}$$

Similar to the true positive rate, the true negative rate is the percentage of actual bad that were accurately categorized as such:

$$\mathrm{TN} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{2}$$

One of the most often used metrics in the field of accounting and finance, specifically for applications involving credit rating, is the Percentage Correctly Classified. The PCC rate calculates the percentage of cases in a given data set that are correctly classified as having excellent credit and having bad credit. The PCC rate is an important factor to consider while assessing the proposed scoring models' capacity for classification.

$$PCC = \frac{TP+TN}{TP+FP+FN+TN} \tag{3}$$

To assess the accuracy of the model's predictions of loan default, the Receiver Operating Characteristic (ROC) curve, a standard classification statistic, is used. The probability of binary outcomes, which are typical in our situation for default and non-default, can be predicted using the ROC curve. The ROC curve compares the ratio of false positives to genuine positives (Osei et al. 2021).

The advantage of the ROC curve is that different thresholds or modeling techniques can be compared using measurements of the area under the curve (AUC), with a greater AUC signifying a better model. The movement along a line indicates a change in the threshold used to classify a positive instance. Each line on the plot represents the curve for a single model.

The threshold is 0 (upper right) and 1 (lower left) (lower left). The AUC, which ranges from 0 to 1 with a good model scoring higher, is the area under the ROC curve. A ROC AUC of 0.5 results from a model's random predictions. Since Sensitivity and (1 - Specificity) are plotted, the ROC Curve logically plots between True Positive Rate and False Positive Rate.

MODEL RESULT ANALYSIS AND DISCUSSION

The model performance was assessed using the metrics that were presented in Chapter 3. It might be difficult and subjective to agree on a single criterion for evaluating performance, depending on the nature of the activity at hand. The quality of a classification model can be evaluated via inspection of the confusion matrix.

After that, a 70/30 split was performed randomly across the entire data set to create a training and test set. Sampling methods such as Tomek-links, kNN, and SMOTE can be used on imbalanced datasets like the one used in this research. The study, however, opted to simply employ a random over-sampling strategy for the response variable due to time and technical limitations. The test data set was used to evaluate how well the models performed after being trained on the training set to make predictions.

A confusion matrix is a table that is able to compare the model predicted and actual classes from the labeled data within the validation set. Hence, the confusion matrices for each of our ensemble classifiers are shown in Figure (2).

The ideal evaluation model would actually be a profit function, which is a function of recall and precision and would need to be optimized. Trade between the TP (profit) and the FP (cost), both of which are captured by the F-measure or the accuracy recall AUC, would be used to estimate the profit. The F-measure, however, is the metric picked for this study's evaluation of the models. The model with the best F measure can then be modified and verified in an effort to produce improved evaluation metrics and predictions.

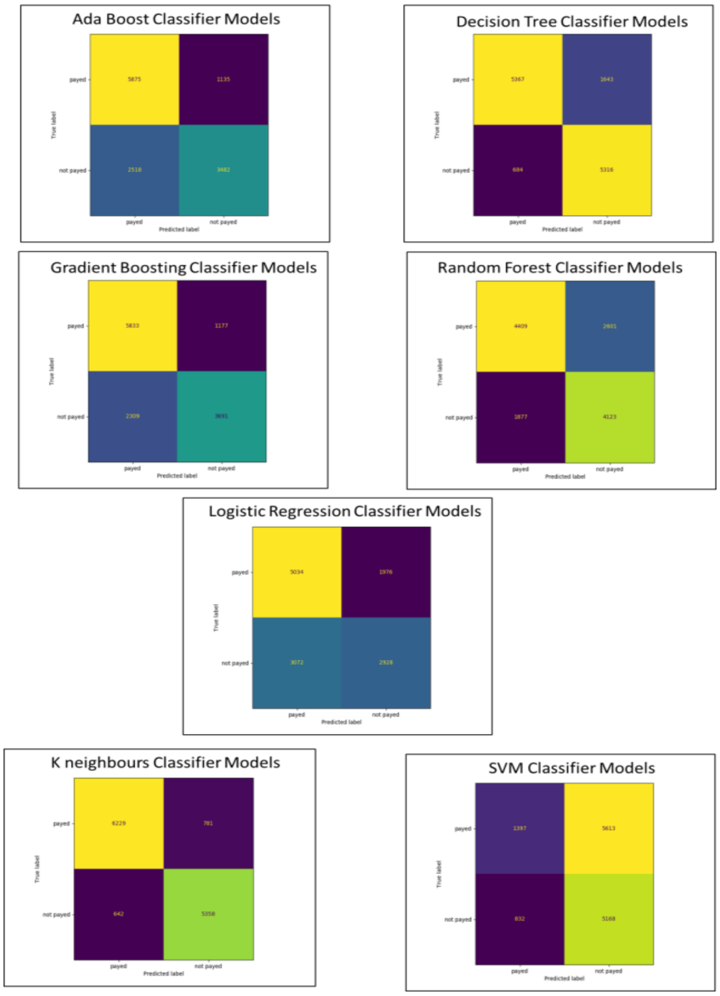

FIGURE 2. The Confusion Matrix for diffrent Classifier Models

TABLE 3. The Summary of Results of Each Model

Model	Accuracy	Recall	Precision	Roc-AUC
Logistic Regression	61%	48%	60%	64%
K neighbours Classifier	70%	78%	65%	78%
SVM Classifier	55%	47%	64%	57%
Decision Tree Classifier	88%	95%	81%	86%
Random Forest Classifier	93%	95%	91%	97%
Ada Boost Classifier	71%	57%	75%	78%
Gradient Boosting Classifier	73%	60%	76%	79%

The results in Table 3 below demonstrate how well the models performed using the measures after the best model for each class was chosen.

Using the results table above as a guide, We conducted an estimation of the range of values for the measures that could potentially be employed to evaluate our models. The Random Forest model, followed by the Decision Tree, has the best accuracy score, while the SVM Classifier has the lowest precision score. On the other hand, the Random Forest and Decision Tree both have the greatest recall scores, whereas Logistic Regression has the lowest. In contrast, the Random Forest has the highest precision score, followed by the Decision Tree, and the Logistic Regression has the lowest.

Figure (3) show the ROC curve of each model. The ROC curve for Random Forest has a convex circle shape, which is indicative of reduced rates of false negative and false positive errors when compared to other curves. In other words, for each given value of sensitivity and specificity, Random Forest performs optimally. In addition, there is little to no difference in performance between Ada Boost, K neighbors, and Gradient Boosting. Random Forest and Decision Tree are the most promising alternatives, yet it is impossible to discern a preferred individual technique from the curves. The skewed ROC curves observed in Logistic Regression, K neighbors SVM, Ada Boost, and Gradient Boosting indicate that the increased specificity resulted in a markedly reduced sensitivity.

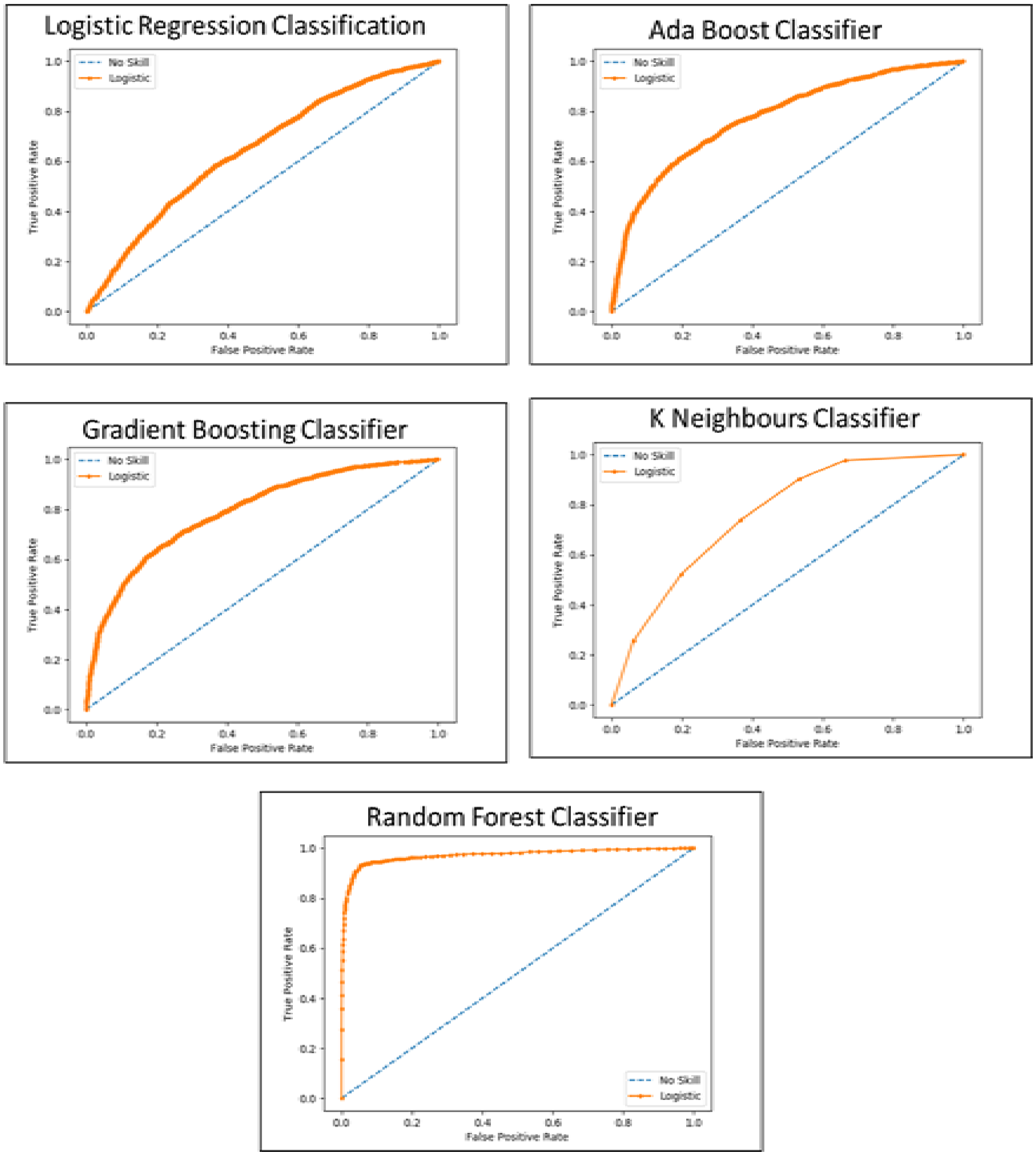

FIGURE 3. The ROC Curve for modles Classification

Choosing the kind of false that the Bank can tolerate while dealing with credit risk is essential. False positives will force us to turn away consumers who would otherwise be profitable clients because the models incorrectly labeled them as the Bank's worst customers, which Precision Score determines. An increase in the Bank's risk due to false negative results consumers as low-risk, but in reality, they would be more likely to default and cause losses for the business, as determined by the Recall score.

Figure (4) displays the Precision-Recall (PR) curves for each model, and the AUC values for each model are all greater than 0.5, indicating that the models performed well.

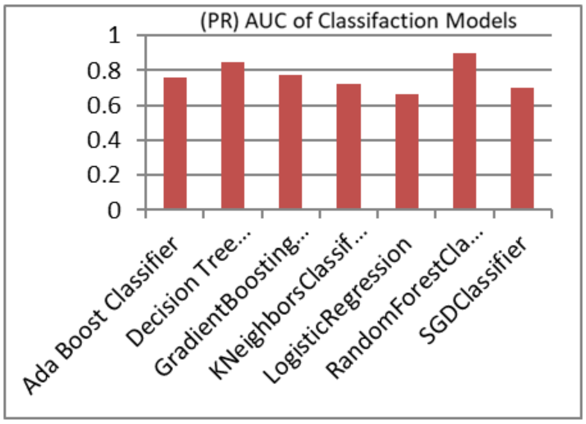

FIGURE 4. The PR(AUC) of Different Classification Models

CONCLUSION AND FUTURE WORKS

Since identifying which customers pose a high risk is more like a sliding scale than a straightforward binary decision, the present study demonstrates the challenges of credit risk modeling. The outcomes have shown that the challenging aspect of doing binary classification is defining a boundary. The imbalanced nature of the dataset further exacerbated the difficulty of the algorithms' learning.

For the data used in this study, it was demonstrated that when using six data mining classification techniques in this research: K neighbors, SVM, Decision Tree, Random Forest, Ada Boost, and Gradient Boosting. Machine learning models outperformed logistic regression, and the best machine learning model for credit risk estimation was Random Forest. The random forest method can handle unbalanced data with hundreds of variables. It automatically balances data sets with rare classes. The logistic regression model performed poorly, and the dissecting Tree model came in second. In this study, Models were considered efficient if the ability to maximize revenue while minimizing the opportunity cost of false positives and false negatives in order to maximize a company's profitability. A statistic called the ROC curve score, a harmonic mean of Precision and Recall, was used to do this. However, in terms of Accuracy (the proportion of properly classified defaults to all observations), Random Forest performed well, scoring 93%. The score was not used as a performance indicator since Accuracy does not consider the cost of misclassification, as shown by the false positives and negatives.

The study concludes that machine learning models are more effective at estimating credit risk when dealing with unbalanced information, such as credit data sets. However, the models might have performed even better had the dataset's number of features been higher. Additionally, more advanced sampling strategies like SMOTE may have helped to balance out the unbalanced data set and boost performance. Consequently, this demonstrates that the findings are not restricted to any one particular bank and may be applied globally to the forecasting of early instances of corporate insolvency.

Investigate various dataset pre-processing techniques, such as feature selection or data-filtering techniques, and ascertain the potential effects on the outcomes. Try to utilize a filtering condensing strategy rather than pure filtering, which will remove both outlier items and non-informative entries that could negatively affect the training process.

ACKNOWLEDGEMENT

This research has received no external funding.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Ala'raj, M., & Abbod, M. 2015. A systematic credit scoring model based on heterogeneous classifier ensembles. *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA):* 1–7.

Anderson, R. 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

Aniceto, M. C., Barboza, F., & Kimura, H. 2020. Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal* 6(1): 1–14.

Bastos, J. 2007. *Credit scoring with boosted decision trees*.

Boughaci, D., & Alkhawaldeh, A. A. 2018. Three local search-based methods for feature selection in credit scoring. *Vietnam Journal of Computer Science* 5(2): 107–121.

Bunker, R. P., Zhang, W., & Naeem, M. A. 2016. Improving a credit scoring model by incorporating bank statement derived features. *ArXiv Preprint ArXiv:1611.00252*.

Condori-Alejo, H. I., Aceituno-Rojo, M. R., & Alzamora, G. S. 2021. Rural micro credit assessment using machine learning in a Peruvian microfinance institution. *Procedia Computer Science* 187: 408–413.

Dastile, X., Celik, T., & Potsane, M. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91: 106263.

Durand, D. 1941. *Risk Elements in Consumer Installment Financing*. National Bureau of Economic Research, New York.

Leo, M., Sharma, S., & Maddulety, K. 2019. Machine learning in banking risk management: A literature review. *Risks 7*(1): 29.

Moradi, S., & Mokhatab Rafiei, F. 2019. A dynamic credit risk assessment model with data mining techniques: Evidence from Iranian banks. *Financial Innovation* 5(1): 1–27.

Mukid, M. A., Widiharih, T., Rusgiyono, A., & Prahutama, A. 2018. Credit scoring analysis using weighted k nearest neighbor. *Journal of Physics: Conference Series* 1025(1): 12114.

Nyangena, B. O. 2019. *Consumer Credit Risk Modelling Using Machine Learning Algorithms: A Comparative Approach*. Strathmore University.

Osei, S., Mpinda, B. N., Sadefo-Kamdem, J., & Fadugba, J. 2021. *Accuracies of some Learning or Scoring Models for Credit Risk Measurement*.

Ozgur, O., Karagol, E. T., & Ozbugday, F. C. 2021. Machine learning approach to drivers of bank lending: evidence from an emerging economy. *Financial Innovation 7*(1): 1–29.

Pławiak, P., Abdar, M., Pławiak, J., Makarenkov, V., & Acharya, U. R. (2020). DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring. *Information Sciences* 516: 401–418.

Tang, L., Cai, F., & Ouyang, Y. 2019. Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China. *Technological Forecasting and Social Change* 144: 563–572.

Yeh, I.-C., & Lien, C. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36(2): 2473–2480.

Zhang, Y., & Wang, J. 2016. K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. *International Journal of Forecasting* 32(3): 1074–1080.

Zhou, X., Cheng, S., Zhu, M., Guo, C., Zhou, S., Xu, P., Xue, Z., & Zhang, W. 2018. A state of the art survey of data mining-based fraud detection and credit scoring. *MATEC Web of Conferences* 189: 3002.