Evaluating Various Machine Learning Techniques in Credit Risk Area

Dongtan Li *

Robert B. Willumstad School of Business, Adelphi University, Garden City, NY, United States

*Corresponding author: dongtanli@mail.adelphi.edu

Abstract. Implementing machine learning techniques to credit scoring is a popular method, which is widely used by many financial institutions and banks at present. As the fast development of machine learning tools, these technologies could provide people more accurate predictions and help enterprises avoid future risk. A supervised machine learning technique is utilized in this research as the classification approach. In this experiment, several machine learning algorithms will be compared in order to present the performance by evaluating the type of credit risk. The data is about assessing customers of a German banking systems from the UCI Machine Learning Repository, which contains 5000 instances and 21 attributes. The final result of this research shows the comparison of 12 scenarios among different combinations of balancing methods, feature selection methods, and predictive algorithms, which finally presents that the collection of Adaptive Synthetic, Boruta and k-Nearest Neighbor receives the highest accuracy score.

Keywords: Credit risk; Supervised machine learning; Adaptive Synthetic; Boruta; k-Nearest Neighbor.

1. Introduction

With the impressive evolution of economy and computer science, data related to financial and credit risks are increasingly collected and stored in digital. Many financial institutions are trying to use different machine learning algorithms to gradually establish and improve their own credit risk prediction system [1]. There is no doubt that a sound credit prediction system can help society, enterprises, and individuals develop more rapidly and safely.

In the era of big data, it is very important for the government to establish a perfect credit system. From the perspective of society, a highly honest society will bring people more sense of security to stabilize the society. The government should bear the responsibility to gather all the information from different Internet platforms to form a complete credit investigation system. This credit investigation system can break the barriers of industry information to enhance the data sharing of social credit information. In addition, the government should enact some laws to protect enterprise and personal credit data to prevent these data being abused and leaked. From the perspective of enterprises, it is necessary to apply machine learning methods to create constructive models. The credit risk prediction model can accurately distinguish good customers and bad customers, which can greatly reduce the risk of credit loan loss of enterprises. Especially for banks, most of the main business income of banks comes from charging interest from borrowers [2]. Therefore, credit risk management is the core point of bank operation management. To some extent, banks can control credit risk well in order to reduce losses, which can help themselves and more enterprises develop. From a personal point of view, the social credit system is also helping people establish a sense of credit step by step. Therefore, more and more people realize that credit score has become the most important determinant to judge whether a person is reliable or not. Nowadays, many people always think about how to improve personal credit score.

2. Literature Review

By reviewing recent studies about credit risk assessment, researchers used different machine learning algorithms to examine the credit records from different banks to predict credit risks. Finally, their research goal is to find a more accurate prediction model to classify credit risks.

Recently, a comparative evaluation of several ML classifiers' performance on credit scoring systems was presented. In this experiment, Pandey et al [2] chose Decision Tree, Baysian Classifier, k-Means, Naïve-Bayes Classifier, Support Vector Machine, Multilayer Perceptron, Artificial Neural Network, k-Nearest Neighbor, and Extreme Learning Machine as their ML techniques. The ensemble of classifiers was also applied to the models since combined classifiers usually had a higher performance than single classifier. In their results, they concluded that Extreme Learning Machine algorithms presented the best accuracies in both German and Australian dataset.

In the study of comparing the difference between supervised and unsupervised learnings, Wang et al. [3] concluded the performance of credit scoring algorithms by combining various supervised and unsupervised machine learning tools in two different stages. The experiment was based on three credit datasets which were divided into four groups, such as individual models and clustering models. During the processing phase, they used 7 supervised and 2 unsupervised machine learning approaches. The supervised learning tools are Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Decision Trees, Support Vector Machines, k-Nearest Neighbors, and Artificial Neural Network. The unsupervised models are k-Means and Kohonen's Self-organizing Maps. The results proved that the integration showed a great improvement on the performance of credit prediction models.

The third research, named Credit Risk Assessment using Machine Learning Techniques, focused on the accuracy of credit evaluation by utilizing different machine learning algorithms. Aithal, V and Jathanna, R. D [4] aimed to discover the best analysis model in order to determine credibility. The methodologies, which they selected, are Logistic Regression, Random Forest, Neural Network, Support Vector Network, Naïve Bayes Classifier, and Classification and Regression Trees. According to their analysis, Random Forest algorithms showed higher accuracy than other approaches on credit risk evaluation.

As Chen [5] mentioned that because of the inaccuracy and inefficiency of traditional credit estimation methods, authors intended to build effective models to discriminate low-credit people from high-credit people in unbalanced data. In this condition, the paper tested four tree-based machine learning algorithms to the database, which are Random Forest, Bagging, Decision Tree, and Adaboost. Although accuracy of their models was not enough to meet the requirements, this research still presented the feasibility of ML algorithms.

Twala, B [6] was to investigate how multiple classifier implementation attributes to credit risk prediction in the research. To examine the impact of various ML and SPR methods' combinations, authors applied 3 supervised learning algorithms and 2 statistical pattern recognition approaches. There were Logistic Discrimination, k-Nearest Neighbor Classifier, Naïve Bayes Classifier, Decision Trees, and Artificial Neural Networks. The experimental outcomes provided clear evidence that ensemble of classifiers models could increase accuracy of credit risk prediction.

In Geraldo-Campos et al [7] research, they aimed to examine the credit risk status after COVID-19 by applying machine learning algorithms in the Reactive Peru Program. Four independent variables with over 500 thousand observations were analyzed by using Lasso and Ridge regression models. Their results indicated that Lasso regression presented better predictions on credit risk determination in this program.

Bussmann et al [8] collected credit scores from 15000 small and medium enterprises which were supplied by European External Credit Assessment Institution. Researchers tried to use different features of borrowers to predict and explain their credit status. In this study, they chose Extreme Gradient Boost model (XGBoost) to test their data. Compared to the baseline model Logistic regression, XGBoost increased the final performance from 0.81 to 0.93. Therefore, the result is good enough to predict customers' future behaviors.

3. Data Exploration

Professor Dr. Hans Hofmann contributed the data to UCI Machine Learning Repository on November 17, 1994. The dataset from German Bank System described the basic credit information of each customer, which includes 21 independent features and 1 dependent variable, Risk. There is total 5000 observations. The main purpose of analyzing this data is to predict "risk" or "no risk" based on different characteristics. Moreover, Table 1 shows that there are 7 numerical variables and 14 categorical variables. In this case, different analysis algorithms will be applied to this dataset to determine whether a customer's credit risk is good or bad.

Table 1. Characteristics of Each Customer

NUMBER	FEATURES	CLASS
1	Customer ID	Categorical
2	CheckingStatus	Categorical
3	LoanDuration	Numerical
4	CreditHistory	Categorical
5	LoanPurpose	Categorical
6	LoanAmount	Numerical
7	ExistingSavings	Categorical
8	EmploymentDuration	Categorical
9	InstallmentPercent	Numerical
10	Sex	Categorical
11	OthersOnLoan	Categorical
12	Current Residence Duration	Numerical
13	OwnsProperty	Categorical
14	Age	Numerical
15	InstallmentPlans	Categorical
16	Housing	Categorical
17	ExistingCreditsCount	Numerical
18	Job	Categorical
19	Dependents	Numerical
20	Telephone	Categorical
21	Foreign Workers	Categorical
22	Risk	Categorical

4. Methodology

4.1 Balancing Methods

4.1.1 Borderline SMOTE

Borderline SMOTE is an advanced condition of Synthetic Minority Over-sampling Technique approach. Nowadays, balancing data is an indispensable step in most machine learning studies. The prediction on the imbalanced data always lacks the accuracy. For instance, considering a binary categorizing problem, the dependent variable results in "Yes" or "No" among ten thousand observations. However, there are only 10 outputs of "No". The prediction on "No" will become inaccurate, since the sample size is too small. Normally, SMOTE creates new observations from the distance between points in the minority class and their nearest neighbors [9]. It may decrease the predictive accuracy when there are some minority's points outlying and appearing in the majority area. BLSMOTE solves this issue effectively. It could distinguish the noise minority's points while their neighbors are from the majority class, and it uses border points to generate new data.

4.1.2 ADASYN

ADASYN (Adaptive Synthetic) was firstly introduced by He et al [10] and now is widely used in most situations of balancing data. For each observation in the minority class, ADASYN firstly calculates the impurity of the neighbors. Higher majority neighbors result in higher impurity ratios. In this condition, it will synthetize the new points around the observations with higher impurity ratios. Compared to the Borderline SMOTE or SMOTE, this method is much softer and more adaptive on the results.

4.2 Predictive Models

4.2.1 Logistic Regression

Logistic regression is one of the most popular analysis models in area of credit evaluation and has been widely used in binary classification problems. It transforms the linear regression surface to logistic style. According to Figure 1, the upper surface is approaching to 1 and lower surface is close to 0. Therefore, the output of a binary problem always ranges from zero to one. The basic equation of a logistic regression is shown below:

$$y = \frac{1}{1+e^{-\sum b_i x_i}} \tag{1}$$

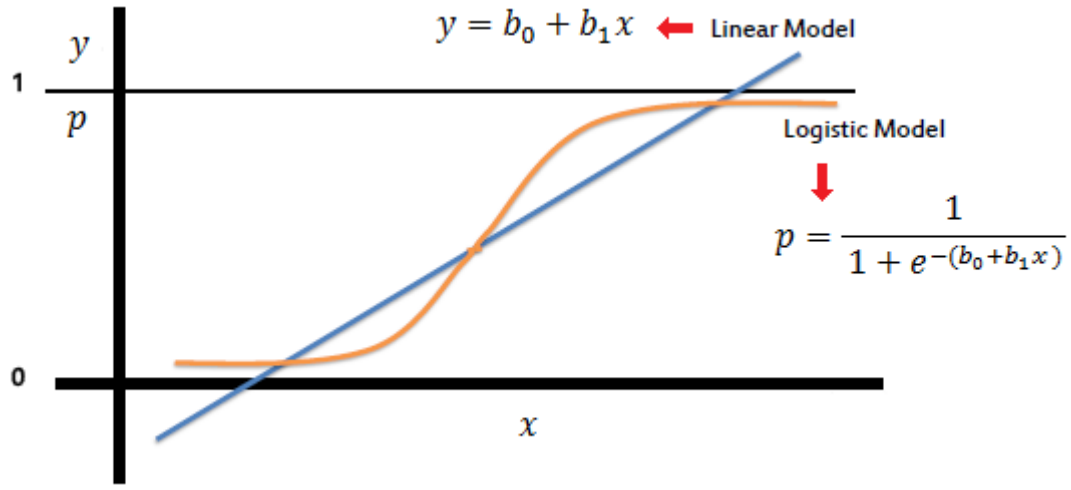

Fig. 1 Logistic Model

4.2.2 Random Forest

Random Forest is the advanced model of decision trees which was firstly created in 1995 and developed by Leo Breiman in 2006[6]. As Figure 2 shown, a random forest is a collection of several strong models. To build a good Random Forest model, only a randomly chosen subset of variables or records is used firstly for each tree and/or for each split iteration within a tree. In other words, it is necessary to build strong and randomness trees at the beginning. Each model or tree will be created independently and have a low correlation. Secondly, all the results are combined whether they are averaging or voting. Since these trees are separated from each other, they will be protected from their individual issues. The characteristic of uncorrelated outcomes will increase robustness, stability, and generalization.

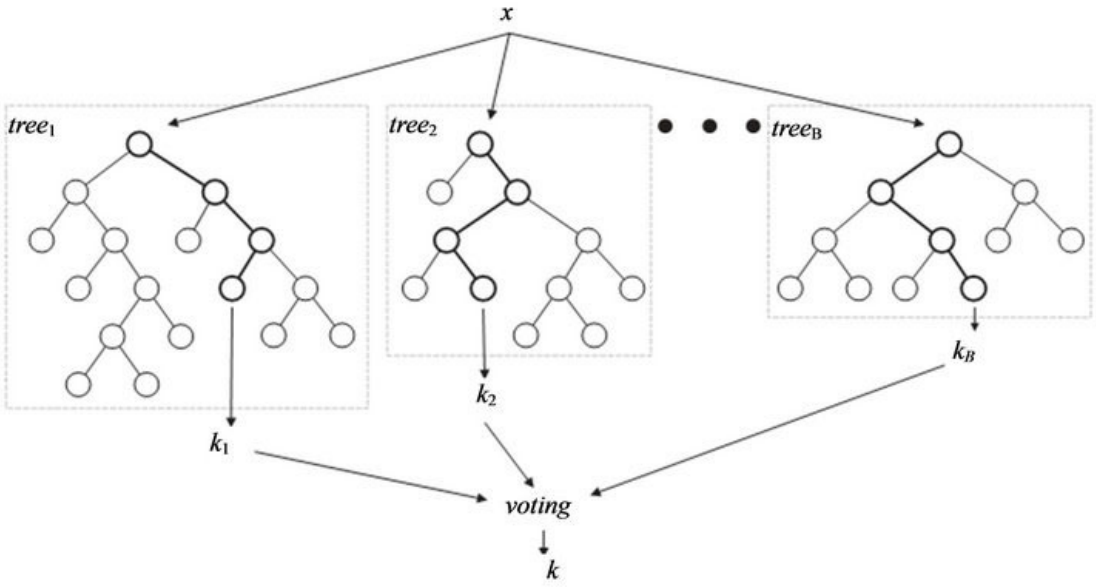

Fig. 2 Random Forest

4.2.3 k Nearest Neighbors

k Nearest Neighbors is a distance-measuring algorithm under the supervised learning. k Nearest Neighbors finds the shortest distance from the target points to the labeled samples. Euclidean distance is one of the distance measuring approaches, which is widely used in k Nearest Neighbors to determine the separation. The formula of Euclidean distance in 2 dimensions is presented as below:

$$D = [|x_{11} - x_{21}|^2 + |x_{12} - x_{22}|^2]^{\frac{1}{2}} \quad (2)$$

However, in most conditions, the studies are multi-dimensional problems. In general n dimension, its equation is presented as below:

$$D = \left[\sum_{i=1}^{n} |x_{1i} - x_{2i}|^2\right]^{\frac{1}{2}} \quad (3)$$

Usually, k Nearest Neighbors works in the following three steps: firstly, it should point in x space for which they want to predict y; secondly, it should find the nearest neighbors around the points selected in x space. The prediction for y is the average of the distance among each nearest neighbor in x space. k Nearest Neighbors is also a popular model in constructing credit scoring systems.

4.3 Data Preprocessing

At the beginning of data preparation, the different types of variables were shown above in order to determine the encoding methods. After removing "Customers ID", there are 20 independent variables and 1 dependent variable -- Risk. Fortunately, the data has no missing value. And the size of "no" class in Foreign Worker variable is too small to provide contributions for the final predictions, which means that Foreign Worker has a near-zero variance. Among these variables, 14 of them are character types and 7 of them are integer types.

4.3.1 Ordinal Encoding

Ordinal encoding is a method to transfer the categorical characters to digits from 1 to n. In this project, it was applied to 5 variables, which are Employment Duration, Sex, Telephone, Foreign Worker, and Risk. The reason is that the observations of employment duration have a natural order and rest of them are binary variables that could be transferred to 1 and 0.

4.3.2 One-Hot Encoding

One hot encoding is one way to create new variables by expanding dimensionality immediately. For example, [A, B, C, D] is assigned to x1, x2, x3, x4 with binary values, so the record with value "B" is [0, 1, 0, 0]. In this experiment, one-hot encoding was implemented to the rest categorical

variables. Moreover, dependent variable Risk was moved to the first column. In the procedure of data preparation, two analysis approaches were used to improve quality of data preparation.

The first method is to check the outliers of the variables. The data have three continuous variables, which are LoanDuration, LoanAmount, and Age. In order to determine whether they are noisy data or not, Box Plot and Cook's distance method were necessary to detect outliers. The box plots of each variable shows that there are some distinct outliers. Cook's Distance is an estimate of influential data points. In addition, Figure 3 generated by Cook's distance method present that there are several points above the middle line, which means that the output also has outliers. Therefore, noisy data binning was applied to these three variables. Four levels were created for LoanDuration and LoanAmount, respectively. Age was divided into 6 levels with the same interval. Finally, there are 45 new columns in the dataset.

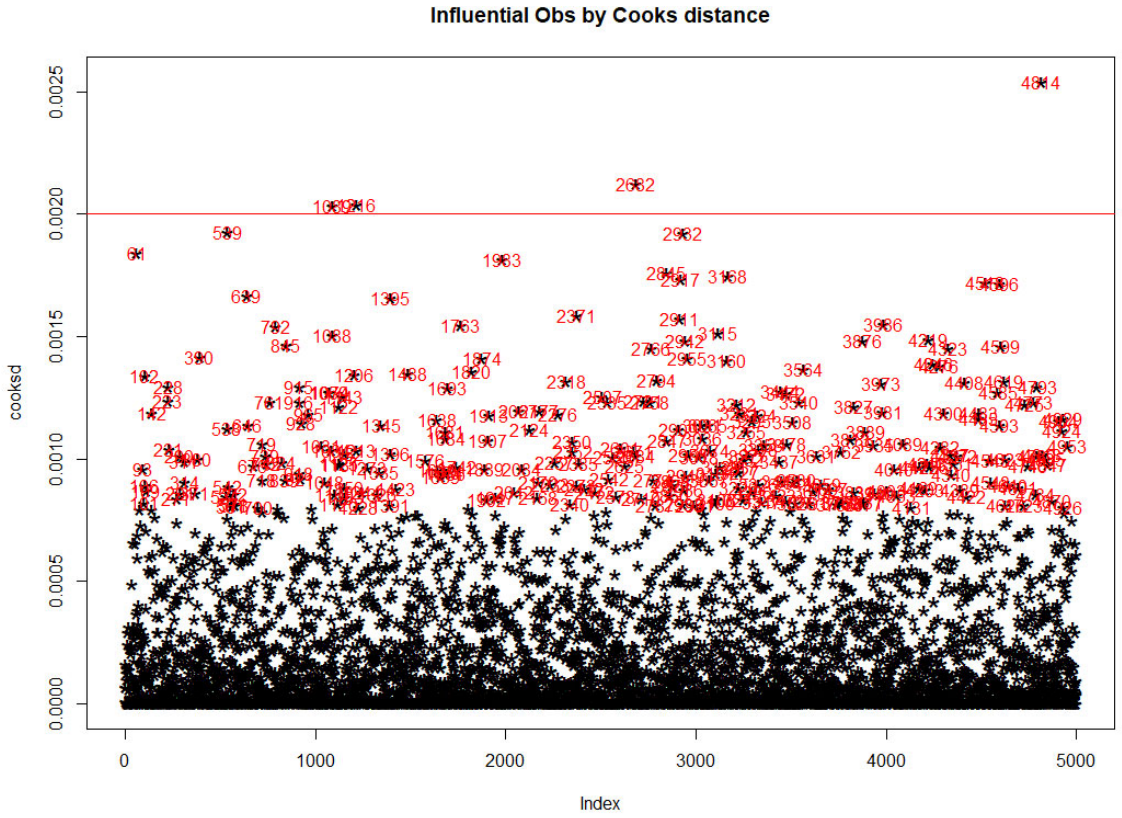

Fig. 3 Cook's Distance

5. Training and testing

The data was split into 30% testing set and 70% training set. However, the data was imbalanced and balancing data is necessary for dependent variable "Risk" by using two different methods of "BLSMOTE" and "ADASYN" with a ratio 0.95. A higher ratio could have the smaller difference between two observations, which are "Risk" and "No Risk." After balancing the data, Random Forest and Boruta were chose to find the important features. After this process, the top 50% of 44 variables were selected, which include 22 variables, and exported to new data for further model training. Figure 4 shows the 22 important features filtered by Random Forest; Figure 5 presents the top 22 features selected by Boruta. In Figure 6, boxplot is used to visualize the features' importance. Finally, the processed training dataset would be trained by three predictive models: Random Forest, Logistic Regression, and k-Nearest Neighbors.

```
 [1] "OwnsPropertyreal_estate"              "EmploymentDuration"
 [3] "CheckingStatusless_0"                 "Age"
 [5] "CurrentResidenceDuration"             "OwnsPropertysavings_insurance"
 [7] "InstallmentPercent"                   "LoanDuration"
 [9] "Sex"                                  "LoanAmount"
[11] "OthersOnLoannone"                     "ExistingCreditsCount"
[13] "CheckingStatusno_checking"            "ExistingSavingsless_100"
[15] "Telephone"                            "Housingown"
[17] "CreditHistorycredits_paid_to_date"    "CreditHistoryprior_payments_delayed"
[19] "LoanPurposecar_used"                  "LoanPurposefurniture"
[21] "ExistingSavings500_to_1000"           "Jobskilled"
```

Fig. 4 Top 22 important features by Random Forest

```
 [1] "OwnsPropertyreal_estate"              "EmploymentDuration"              "CheckingStatusless_0"
 [4] "OwnsPropertysavings_insurance"        "OthersOnLoannone"                "Sex"
 [7] "Age"                                  "CurrentResidenceDuration"        "LoanDuration"
[10] "Housingrent"                          "ExistingSavingsless_100"         "CheckingStatusno_checking"
[13] "Housingown"                           "ExistingCreditsCount"            "InstallmentPercent"
[16] "CreditHistorycredits_paid_to_date"    "Telephone"                       "Dependents"
[19] "LoanAmount"                           "OwnsPropertyunknown"             "LoanPurposecar_new"
[22] "InstallmentPlansstores"
```

Fig. 5 Top 22 important features by Boruta

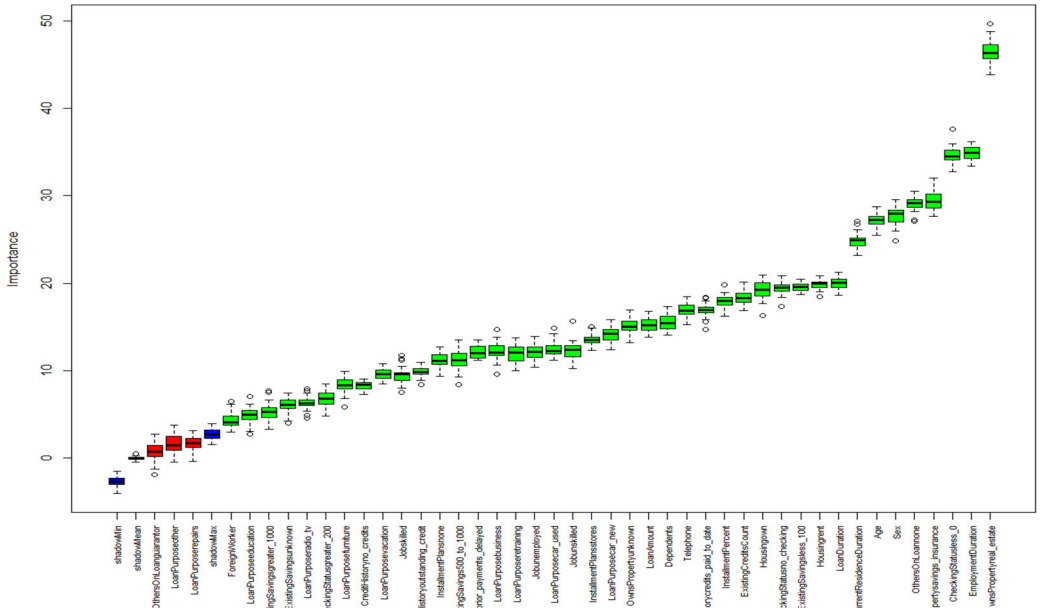

Fig. 6 Boxplot for Features' Importance

6. Analysis and Outputs

In order to test the influence of different feature selection and balancing methods, two feature selection method, two resampling techniques, and three analytical methods were applied to data. The approaches for selecting important features are Random Forest and Boruta. There are some differences between the outputs of these two methods. For instance, "Job Unemployed" was only included in Random Forest extraction. The balancing algorithms are "BLSMOTE" and "ADASYN". The utilized predictive models are Random Forest, Logistic Regression, and k-Nearest Neighbors. According to the table 2, the combinations with the k-Nearest Neighbors method obtained higher scores in all evaluation metrics than other scenarios. The final performance is presented in Sensitivity, Specificity, Precision, G mean, Accuracy, and AUC.

Because G Mean includes both sensitivity and specificity, the group with the largest number of G Mean would be considered as the best model, which contains Boruta (Feature Selection), ADASYN (Balancing), and kNN (Training Model)

Table 2. Scenarios Outputs

Scenarios	Sensitivity	Specificity	Precision	G Mean	Accuracy	AUC
RF+BLSMOTE+RF	0.8139	0.6034	0.7936	0.7008	0.7407	0.7137
RF+BLSMOTE+Boruta	0.8209	0.6157	0.7996	0.7109	0.7493	0.7237
RF+ADASYN+RF	0.8247	0.6119	0.7926	0.7104	0.7487	0.7263
RF+ADASYN+Boruta	0.8245	0.6304	0.8106	0.7209	0.758	0.7312
LR+BLSMOTE+RF	0.8132	0.6356	0.8245	0.7189	0.756	0.7211
LR+BLSMOTE+Boruta	0.8123	0.6394	0.8285	0.7207	0.7573	0.7211
LR+ADASYN+RF	0.8183	0.6582	0.8395	0.7339	0.768	0.7316
LR+ADASYN+Boruta	0.8151	0.6552	0.8395	0.7308	0.7653	0.7276
kNN+BLSMOTE+RF	0.9766	0.8482	0.9152	0.9101	0.9287	0.9355
kNN+BLSMOTE+Boruta	0.9781	0.8815	0.9362	0.9286	0.9433	0.947
kNN+ADASYN+RF	0.9843	0.8877	0.9392	0.9347	0.9493	0.9545
kNN+ADASYN+Boruta	0.9813	0.8937	0.9432	0.9365	0.95	0.9535

7. Conclusion

These are the features that have the greatest impact on the response variable, so banks should pay more attention to these. For example, “OwnsProperty” shows that the credit risk of customers who own real estate is extremely low, and banks can lend to these customers with confidence. For “CheckingStatusless_0” and “OthersOnLoannone”, these also have a significant impact on the predictor variable "Risk". If a customer does not have a checking account, this will increase the credit risk. If he has no other debts, this will also significantly reduce his credit risk. In addition, when assessing credit risk, banks should also pay attention to the customer's "Sex", "Age", "EmploymentDuration", "CurrentresidenceDuration", etc. By knowing the sex, age, duration of employment, and length of residence in the area, the German Bank can better predict the credit risk of the customer.

However, there are some limitations in this project. First, the dataset is from 1994, which means that the source of this data is too old. Second, most of the observations in this data are foreign workers. In future data, more native observations could be expected so that credit risk can be predicted more accurately. Furthermore, more models and more performance measurement could be applied in the future research.

References

[1] Shi, S., Tse, R., Luo, W., D’Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. Neural Computing & Applications, 34(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

[2] Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., &Dehuri, S. (2017). Credit risk analysis using machine learning classifiers. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 1850-1854. https://doi.org/10.1109/ICECDS.2017.8389769

[3] Wang, B., Ning, L. J., & Kong, Y. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications, 128, 301-315. https://doi.org/10.1016/j.eswa.2019.02.033

[4] Aithal, V., &Jathanna, R. D. (2019, November). Credit risk assessment using machine learning techniques. Manipal academy of higher education. https://manipal.pure.elsevier.com/en/publications/credit-risk-assessment-using-machine-learning-techniques

[5] Chen, Z. (2018, November 19). The application of tree-based model to unbalanced German credit data analysis. Edp sciences. https://www.matec-conferences.org/articles/matecconf/abs/2018/91/matecconf_eitce2018_01005/matecconf_eitce2018_01005.html

[6] Twala, B. (2010). Multiple classifier application to credit risk assessment. Expert Systems with Applications, 37(4), 3326-3336. https://doi.org/10.1016/j.eswa.2009.10.018

[7] Geraldo-Campos, L., Soria, J. J., & Pando-Ezcurra, T. (2022). Machine Learning for Credit Risk in the Reactive Peru Program: A Comparison of the Lasso and Ridge Regression Models. Economies, 10(8), 188. https://doi.org/10.3390/economies10080188

[8] Bussmann, N., Giudici, P., Marinelli, D., &Papenbrock, J. (2020). Explainable Machine Learning in Credit Risk Management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[9] Chawla, Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. The Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[10] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic samplingapproach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IJCNN'08), pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.