The Comparisons of the Machine Learning Models in Credit Card Default under Imbalance and Multi-features Dataset

Zhongtian Yu[a]
School of Accounting and Finance, The Hong Kong Polytechnic University, Hong Kong, China

Keywords: Chinese Consumption Credit Card, Machine Learning, Imbalance Data.

Abstract: Affected by the novel coronavirus pneumonia, the global financial market has suffered from a terrible crisis, so the risk tolerance of banks around the world is greatly weakened, which requires the improvement of risk management in banks. The development of machine learning makes programming more convenient and prediction more accurate. In terms of risk management, the introduction of machine learning models enables banks to more accurately predict the potential risks, providing more opportunities to avoid them. China is the fastest recovery country under COVID-19, but previous studies are lack of analysis data from the Bank of China. Therefore, the paper processes the data from the Bank of China to train five models (the support vector machine, decision tree, logistic regression, bagging and random forest) and selects the best model by three standards: effectiveness, efficiency and stability. For achieving the best classification, the paper also tests the optimization effect of feature selection on the five models. In order to ensure the results are fair and universal, the SMOTE is used to solve the problem of data imbalance and grid search is used to obtain the best model parameters, so the influence of parameters on the comparison results between models can be eliminated. Decision tree model performs better considering the complexity and training time and the feature selection does not show improvement in the performance of the tree model in the solution.

1 INTRODUCTION

The coronavirus triggered the global economic disruption in 2020, which is a substantial challenge for policy makers and financial market (Ozili, Arun, 2020). An observable downturn seriously harm the confidence of financial market participants lending to further public anxiety about economy Fetzer, Hensel, Hermle, Roth, 2020).The worries about economic uncertainty and risk lead to the conventical behaviors in investment and the shrink in credit market ,which is bad news to economic recovery and the reestablishment of market confidence .It is of great importance for financial institutions and organizations to balance the risk and investment income rerunning the loan business for revival .The risk management emerged and attended institutions attention in the financial crisis of 2007-2009,more practice and technique show the requirement to the combination of machine learning models and loan departments in banks (Butaru, Chen, Clark, Das, Lo, Siddique, 2016). Machine learning enable computers to learning the patterns of data and find the inherent regulations for predicting or organizing, which means that artificial intelligence have enough ability to complete some special tasks efficiently replacing the work of human (Samuel, 1959).

Researchers in finance subjects focus on the application of machine learning in Fin-tech area, which may be the revolution of industries and academic research. High-frequency trading and electronic market bring new challenge to the existing theories and emerging opportunity to new theories (Linnenluecke, Chen, Ling, Smith, Zhu 2017). In asset pricing, deep learning is used in portfolio optimization (Heaton, Polson, Witte 2017). In risk management, machine learning can be applied in credit card fraud. The fraud is generally divided in two types: application and behavioral frauds (Bolton, Hand, 2001). Because the deification of fraud is an issue based on the classification scoring between 0 and 1, the SVM (support vector machine) may be a good model (Rtayli, Enneya, 2020). More ensemble models for achieving high accuracy energy in credit card fraud

[a] https://orcid.org/0000-0002-6564-470X

Yu, Z.
The Comparisons of the Machine Learning Models in Credit Card Default under Imbalance and Multi-features Dataset.
DOI: 10.5220/0011355600003440
In *Proceedings of the International Conference on Big Data Economy and Digital Management (BDEDM 2022)*, pages 864-872
ISBN: 978-989-758-593-7

research, like RF (random forest), experiment data indicates that RF perform better compared with LR (logistic regression) and SVM. More methods in modeling and data processing such as hyper-parameters optimization, feature elimination, SMOTE technique, the survival analysis model and grid search has been introduced into the research in risk management (Rtayli, Enneya, 2020) (Li, Li, Li 2019).

expect for credit card fraud deification, the credit card default is another significant domain at the leading edge of innovation in risk management. Machine learning models can support the existing credit scoring method, improving the accuracy of default identification, which can efficient control credit risk without repeated manual operation and expert consulting (Husejinovic Admel, Keco Dino, 2018) (Yang, Zhang, 2018).

The scoring method has been recognized in identification of bad and good loans in 1941, while the method is too academic to be applied in banks (DURAND, 1941). Johnson simplified the method and make it available to practice as underlying regulations in financial institution to select loans candidates (Johnson, 1992). Default rates drop by 50% in some organizations that use the scoring method, which shows the observable performance of scoring (Myers, Forgy 1963).

Classification and regression are two mainstream application in machine leaning, the credit card default is a typical field that classification can be applied to predict credit score based on the financial condition of credit card holder (Sariannidis, Papadakis, Garefalakis, Lemonakis, Kyriaki-Argyro, 2019). the early research that combine the scoring and data analysis model tested the performance of the BDT (Bayesian decision tree), discriminant analysis and linear regression (LC, 2000). Nearly 20 years witnesses the tremendous development of the quantitative analysis method in the domain, machine learning and deep learning innovated data analysis model and statistic model (Leo, Sharma, Maddulety, 2019). Those innovative empirical researches may be generally divided to two categories: the analysis on data and features and the comparison of various models.

In this paper, we selected five models: DT (Decision tree), RF (Random Forest), LR and the bagging decision tree and made a comprehensive standard for selecting the best model. The comparison has considered the influence of the hyperparameters in the models, so the grid-search is introduced to achieve the best performance of each model and find the most excellent model in this kind of application. Meanwhile, the SMOTE (Synthetic Minority Oversampling Technique) and feature importance are introduced in the paper for dealing with issues in the dataset and testing the effect of feature selection in those models.

2 DATA

2.1 General Description of Data

The data in the previous studies are similar to the normal distribution. The data from Taiwan's banks is widely used in the previous research and the default customers account for 22.12% in the data (Sariannidis, Papadakis, Garefalakis, Lemonakis, Kyriaki-Argyro, 2019). By searching the papers about customers' behavior, I found that the data are not consistent with the reality in Chinese mainland. In fact, the credit card introduced in China in the late 1970s, so many Chinese customers have not accepted some exclusive 'early consumption'. Therefore, Chinese, especially for the majority of people living in underdeveloped inland areas, take a cautious approach for their consumption loans and the generally preference reduce the arising of default in Chinese customer credit card (Rong, 2018). Many previous studies did not consider the reality in the selection of dataset, the division of train set and test set. this paper directly uses the data from the Bank of China, does not do the normalization processing, and retains all provided available features. There are 45,985 instances in the credit card dataset. Just to clarify, statues recorded the debts situations of users, 0: 1-29 days past due, 1: 30-59 days past due, 2: 60-89 days past due, 3: 90-119 days past due, 4: 120-149 days past due, 5: Overdue or bad debts, write-offs for more than 150 days, C: paid off that month, X: No loan for the month. All users that have status 2 and above will be recorded as risk. Generally, users in risk should be in 3%, thus I choose users who overdue for more than 60 days as target risk users (Block, Vaaler, 2004). Those samples are marked as '1', else are '0'. There are 667 IDs are identified as in risk, accounting for 1.5% in the dataset.

As the side effect, the processing method leads to too few samples in the default class, which may lead to the neglect of minorities and the overwhelmingly imbalance in the data set. Therefore, the paper uses SMOTE balance method to eliminate the impact of unbalanced data.

2.2 The Description of Feature

The processing of features and the nature of client information influence the basic selection of features. Therefore, the 16 features can be classified to 3 types with different processing method: binary features, continuous features and categorical features.

Table 1: General Structure of Features.

Binary features	Continuous features	Categorical features
Gender	Number of children	Income type
Having a car or not	Annual income	Occupation type
Having properties or not	Age	House type
Having a phone or not	Working years	Education
Having an email or not	Family size	Marriage condition
Having a work phone or not		

The binary features can be straightly processed in '0' and '1'. In gender, the female, for example, is identified as '0' and the male is as '1'.

The continuous features are cut into several groups based on different levels. For example, the number of children is clustered into 3 groups: the family without child, 'ChldNo_0', one child, 'ChldNo_1' and family with more than 2 children,' ChldNo_2More'.

The categorical features are processed by '==' to classify the variables into different groups. The process is just like that the education level are recorded in 3 types: Higher education, 'edutp_Higher education', Incomplete higher, 'edutp_Incomplete higher' and other, 'edutp_Lower secondary, so the computer matches the variables to the 3 labels with heading word, 'edutp', and records '0' or '1' as yes or no in those labels.

2.3 The Analysis of Data

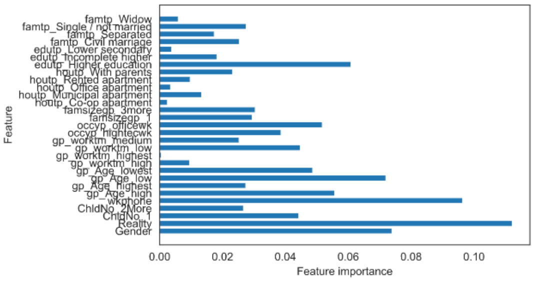

Figure 1: Feature Importance Based on the Features Used in Machine Leaning.

The feature importance analysis is trained in the DT and the similar distribution is known from the test in RF. The maxima figure is about 0.1 and the others are less than the standard, which suggest that there is no feature having decisive role in the model building. Meanwhile, it is conspicuous that the binary feature generally shows higher scores in the '0' and;'1' issue.

2.4 Heatmap

Heatmap is an effective visualization tools in data analysis. It can present a general tend or distribution of dataset. In the heatmap, the darkness of bars represents the performance of each IDs, so the credit card owners with the lighter color, means having better credit record. Also, there are two axes in the image, vertical one is IDs and each one of them represent a card holder and another is Month balance, indicates the months before current time. We can observe most of the map is white, which means majority has good credit records. But there are a few black areas, this one turns from light grey to sudden black, which means he may have some considerable accidents that crushed him. And this longest one is gradually changed from light gray to all black, and has been maintained. It shows that he has been going downhill since then, and he is working hard to make up for it, but the situation is still getting worse. And this black stopped abruptly, indicating that he was unable to recover after the situation deteriorated and his credit card was revoked.

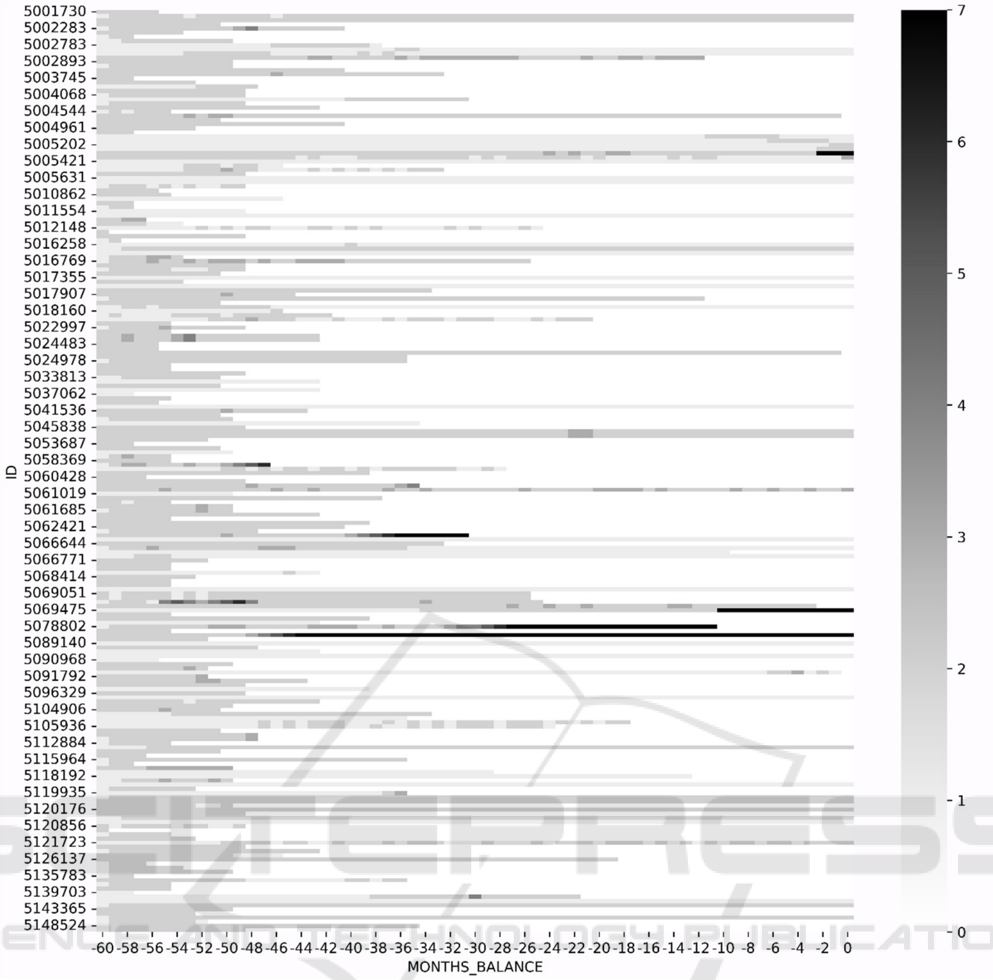

Figure 2: Heatmap based on Id, Months_Balance.

3 METHODOLOGY

3.1 Data Processing Methodology

In the division of training and test set, the SMOTE is used for resolving the minority issue. The data set is not balanced and there are very few people who do not comply with credit, and the machine learning algorithm is likely to ignore the minority class, and thus perform poorly in this class. Because we only have two classes, good and bad, this defect is fatal. So, we used SMOTE to refit the sample. SMOTE can transform an imbalanced dataset to a balanced one by producing arbitrary examples rather than simply oversampling through duplication or replacement (Han, Wang, Mao, 2005). Although, the data processed by SMOTE is in some terms changed in structure, while if the models have high performance in test set under cross validation, the models have ability in adaption of imbalance data with the support of SMOTE.

3.2 Machine Learning Models

I considered the characteristics of different models and reviewed the selection in the previous papers. Therefore, the paper uses five models cover clustering and regression models in the experiment.

3.2.1 The Decision Trees

The model is based on Bayesian optimization methods. “Decision Tree” is a tree-like structure that allows the system to make decision by weighing the possible actions. Bayesian optimization method is used in the model for achieving the weight of each node. The features of data set are the internal nodes and the additional nodes are continuously created from the internal nodes until all instances have been covered by the leaves of the tree (Kotsiantis, Zaharakis, Pintelas 2006).

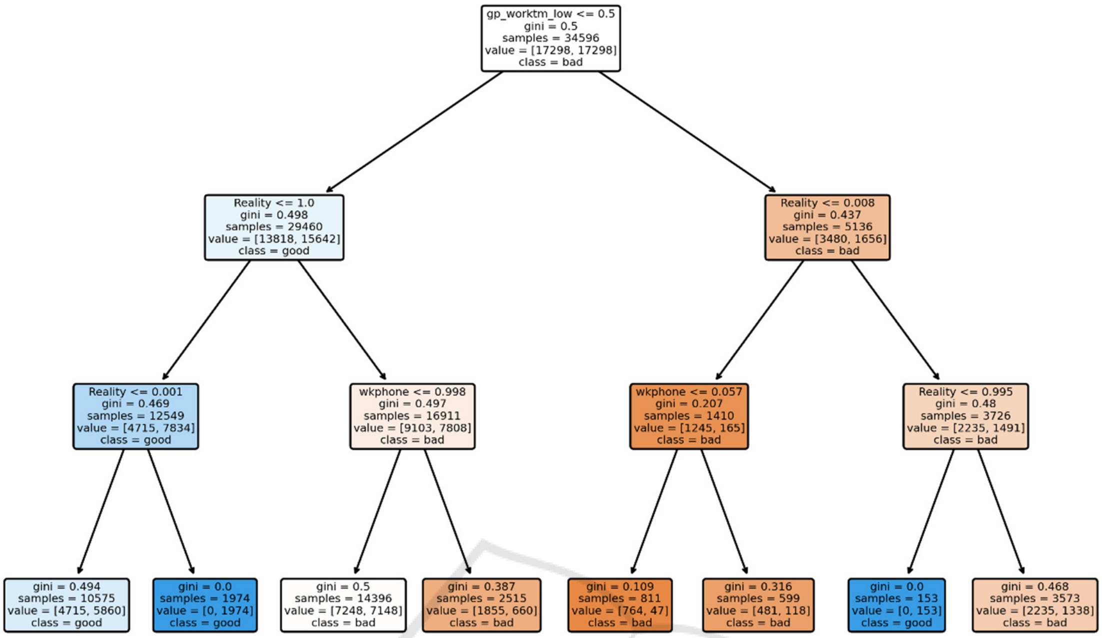

Figure 3: Decision Tree Based on Gini Is Trained by the Data Set of Credit Card Portfolio and Each Node with the Features' Values Is Classed in True and False (Good and Bad Credit Card Owners).

3.2.2 The Bagging

Bagging is formed by a group of decision trees. As mentioned earlier, each node in the decision tree model is given calculated weight, so each tree in the group is given suitable weight. Each tree runs independently to access the best performance and the final outcome of the model is based on the voting of all trees. Therefore, the bootstrap is introduced in the generation of trees to minimize the correlation of each tree.

3.2.3 The Random Forecast

The principle of the random forecast is same as the bagging, but each tree is trained by all features in bagging and by a part of features in random forest. This means that the bagging performs better than random forest when there are just several trees in the modeling. With the increase of trees, the performance of RF shows more obvious improvement than bagging.

In general, RF perform better in sheer data and lots of features.

3.2.4 the Support Vector Machines (SVM)

SVM is a binary linear prediction model. To be specific, the model makes a line that can accurately distinguish two types of points in area or space. In the credit card issue, the model runs in feature space with high dimensions, but the principle of model has not changed. The general principle is to search a large margin decision boundary. The boundary is decided by the distance of the closest points of two sets and has the largest distance to the each of the two points.

3.2.5 Logistic Regression (LR)

LR is based on logistic function, also called the sigmoid function and the curve of LR is an 's-shape' curve. The values of the function is range from 0 to 1 and the curve is generally symmetry about point in 0.5. Therefore, the function is extremely suitable for simple classification. In the classification problem, the Maximum-likelihood estimation, a common learning algorithm is used to search for the best coefficients. The model with the result of training can predicts a value very close to 1 for the default class and a value very close to 0 for the non-default class in credit cards portfolio (Sariannidis, Papadakis, Garefalakis, Lemonakis, Kyriaki-Argyro 2019).

4 RESULT

4.1 Gridsearch for Best Parameters Results

For removing the effect of parameters in model comparison, grid search is applied for searching the best parameters of the models. The further research about feature and model selection is based on the best parameters. The Table2 shows the parameters used for the models and the exact values.

Table 2: The Table Show All Parameters and Their Best Values of the Five Models.

	dt	rf	svm	bag	LogR
C			100		36
n_estimators		25		28	
random_state	13	17		16	1
max_depth	26	24			
gamma			10		

4.2 Five Times Five-Fold Cross Validation and ROC Results

In this paper, 5 times of five-fold cross-validation is used to verify the model established by different data mining methods in best parameter and some scores are accessed from the cross-validation by record and calculation might be also useful for model selection. Meanwhile, the ROC (receiver Operating Characteristic) ratio is introduced to evaluate the effectivity and general performance of the models in another aspects. The detailed result is show in Table3 and Table 4.

Table 3: Accuracy Rates of 5 Times of Five-Fold Cross-Validation Cross Validation.

	1	2	3	4	5
dt	0.90867	0.901632	0.906526	0.907970	0.902471
rf	0.908235	0.901634	0.906526	0.908115	0.902616
svm	0.903631	0.902032	0.903940	0.903792	0.897453
bag	0.907943	0.904067	0.908543	0.908262	0.904071
LogR	0.639568	0.637821	0.635343	0.635018	0.632031

The table includes the mean of five times' score of each fold cross validation as the '1-5' rows' values of 5 folds cross validation.

Table 4: Comparison of Model Classification Effect.

	5_fold fit time	ROC	mean(acc)	std(acc)
dt	0.498136044	0.900391152	0.905453768	0.003215
rf	2.146449327	0.900998112	0.90542531	0.003107
svm	218.6840575	0.896209873	0.902169578	0.002746
bag	9.456052542	0.900256272	0.906577416	0.0023
LogR	1.423333883	0.631710278	0.635956208	0.00288

5_fold fit time is the sum of fit time spend on 5-fold cross-validation (based on the average fit time of 5 times cross-validation). The ROC is a Comprehensive evaluation score. The mean(acc) is equal to the mean in Table3, showing the average accuracy rate of the models. The std(acc) is the standard variances of the accuracy rate show in Table3.

4.3 The Correlation of the Cross Validation of the Models

The accuracy rate accessed has shown above, while the figures are similar. Therefore, the correlation matrix should be introduced to evaluate the performance difference among the models. By visualization, the Fig.4, shows the coefficients directly.

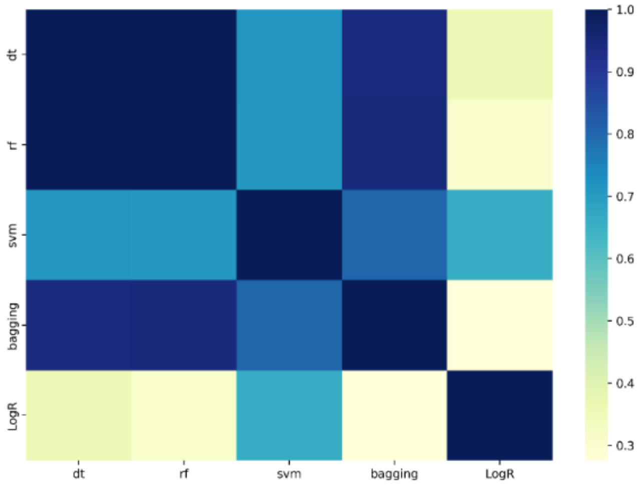

Figure. 4: Correlation of the Cross-Validation Results.

The abbreviations of the models make up the X and Y axes, forming the coordinate. The colors of blocks are defined by the coefficients in matrix formed by Spearman correlation coefficient. The color bar in the right of the picture shows the range of the coefficients (from 0 to 1).

4.4 Feature-Selection Results

All features have used in the paper as the basic research shown above, while the selection of feature is generally useful in previous studies, so the paper tests the effect of feature selection in the models. The chi-square test measures dependence between stochastic variables is recalled for evaluating the feature importance and the TABLE.5. shows the accuracy rate with all features and the top 10% features selecting by the ratios (FIG.1.).

Table 5: Result of Feature Selection.

	LogR	svm	dt	bag	rf
acc	0.63171	0.89621	0.90039	0.90026	0.90100
acc(top10%)	0.63859	0.86950	0.87099	0.87146	0.87294

The acc is the accuracy rate without feature selection. The acc(top10%) is the accuracy rate with feature selection, retaining the features with the top 10% features with the highest feature importance.

5 DISCUSSION

The paper cannot determine which one is the best, because expect for the logistic regression, the other 4 models has the similar performance in accuracy rate and ROC. The logistic regression performs extremely worse than the others, so it is ruled out for the further selection.

The correlation matrix in Fig.4. provides that the variance among the svm, dt, bagging and rf in accuracy rate is extremely little. Therefore, the random forest has the highest accuracy rate and ROC rate, but ‘the best model’ cannot be given to the random forest. This is because the stability is worse than the svm and bagging. Meanwhile, the fit time is almost the 4 times of the dt. Therefore, if the volume of data is not big and the requirement of stability is high, the rf is the best model. However, if the fit time or the stability of outcome is the priority of programming design, the dt and bagging are the best choices respectively.

The feature selection is a common machine learning method for improving accuracy. However, it does not work well in my research. The performance of 4 models performing better are not further improved, but became worse in accuracy. The LogR is the worst model of the 5 models in accuracy, while the accuracy rate increases contrarily.

6 CONCLUSION

The objective of the paper is finding the best model in predicting credit card default. In the processing of reviewing previous studies, two issues emerged. The first is the data set used is same from bank in Taiwan, but the Chinese mainland is significant financial markets, especially at present under the destruction of coronavirus. Therefore, the paper uses the data with the Chinese features. The second is the parameters of the models is assigned without any standards, so the parameters may affect the results. Therefore, the grid search is applied for searching the best parameters of the models and the comparison of the models becomes exacter, because all performance of the models is the best with the best parameters.

The standard for evaluation includes the 5-fold accuracy rate (from 5 times cross validation), the ROC rate, the fit time, the standard variance and mean of accuracy rate. Those ratios can provide comprehensive evaluation in effectiveness, stability and efficiency. According to the standard, the 3 models: the random forest, decision tree and bagging show outstanding performance. The random forest has the highest ROC and the accuracy rate, so the random forest is the best in effectiveness. The decision tree has the similar accuracy and ROC, while the fit time is extremely smaller than the random forest, so the decision tree is the most efficient model. The bagging performs best in the standard variance of the accuracy rates, so the performance of the bagging is steadier than the others.

For accessing the best result, the paper tests the effect of feature selection, while the performance is very bad. The 4 models: the random forest, decision tree, bagging and SVM suffers from obvious fall in accuracy, except for the logistic regression.

The ratios calculated in the paper do not have much academic value, while the comparison results have practical and academic value in some distance. Meanwhile, the thinking of data processing, model training and comparison standard shown above may inspire some later scholars to test the application of the new models in risk management field.

REFERENCES

A.~L.~Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” IBM J. Res. Dev., vol. 3, no. 3, 1959.

D. DURAND, Risk elements in consumer installment financing, vol. 17. New York: National Bureau of Economic Research, 1941.

F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, “Risk and risk management in the credit card industry,” J. Bank. Financ., vol. 72, pp. 218–239, Nov. 2016, doi: 10.1016/j.jbankfin.2016.07.015.

H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3644 LNCS, pp. 878–887, 2005, doi: 10.1007/11538059_91.

J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” Applied Stochastic Models in Business and Industry. 2017, doi: 10.1002/asmb.2209.

J. H. Myers and E. W. Forgy, “The Development of Numerical Credit Evaluation Systems,” J. Am. Stat. Assoc., vol. 58, no. 303, pp. 799–806, 1963, doi: 10.1080/01621459.1963.10500889.

J. Rong, “A study on the relationship between credit card quota and credit card consumption attitude(in Chinese),” Nanjing Univ., p. 67, 2018.

M. K. Linnenluecke, X. Chen, X. Ling, T. Smith, and Y. Zhu, “Research in finance: A review of influential publications and a research agenda,” Pacific Basin Finance Journal, vol. 43. Elsevier B.V., pp. 188–199, Jun. 01, 2017, doi: 10.1016/j.pacfin.2017.04.005.

M. Leo, S. Sharma, and K. Maddulety, “Machine learning in banking risk management: A literature review,” Risks, vol. 7, no. 1, Mar. 2019, doi: 10.3390/risks7010029.

M. Z. Husejinovic Admel, Keco Dino, “Application of Machine Learning Algorithms in Credit Card Default Payment Prediction,” Int. J. Sci. Res., vol. 7, no. 10, pp. 425–426, 2018, doi: 10.15373/22778179#husejinovic.

N. Rtayli and N. Enneya, “Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization,” J. Inf. Secur. Appl., vol. 55, no. September, p. 102596, 2020, doi: 10.1016/j.jisa.2020.102596.

N. Sariannidis, S. Papadakis, A. Garefalakis, C. Lemonakis, and T. Kyriaki-Argyro, “Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques,” Ann. Oper. Res., Nov. 2019, doi: 10.1007/s10479-019-03188-0.

P. K. Ozili and T. Arun, “Spillover of COVID-19: Impact on the Global Economy,” SSRN Electron. J., Mar. 2020, doi: 10.2139/ssrn.3562570.

R. J. Bolton, and D. J. Hand, “Unsupervised Profiling Methods for Fraud Detection,” Proc. Credit Scoring Credit Control VII, 2001.

R. W. Johnson, “Legal, social and economic issues implementing scoring in the US.,” Thomas, L. C \, 1992.

S. A. Block and P. M. Vaaler, “The price of democracy: Sovereign risk ratings, bond spreads and political business cycles in developing countries,” J. Int. Money Financ., vol. 23, no. 6, pp. 917–946, Oct. 2004, doi: 10.1016/j.jimonfin.2004.05.001.

S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: A review of classification and combining techniques,” Artif. Intell. Rev., vol. 26, no. 3, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

S. Yang and H. Zhang, “Comparison of Several Data Mining Methods in Credit Card Default Prediction,” Intell. Inf. Manag., vol. 10, pp. 115–122, 2018, doi: 10.4236/iim.2018.105010.

T. Fetzer, L. Hensel, J. Hermle, and C. Roth, “Coronavirus Perceptions and Economic Anxiety,” Rev. Econ. Stat., pp. 1–36, 2020, doi: 10.1162/rest_a_00946.

T. LC, “A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers,” Int. J. Forecast., vol. 16, p. 149, 2000.

Y. Li, Y. Li, and Y. Li, “What factors are influencing credit card customer’s default behavior in China? A study based on survival analysis,” Phys. A Stat. Mech. its Appl., vol. 526, p. 120861, Jul. 2019, doi: 10.1016/j.physa.2019.04.097.