Should machine learning be applied in credit risk accessment

Yuejia Zhu[1,5,9,†], Jiayi Zhu[2,6,†], Jiaoyuan Ding[3,7,†], Ziyan Li[4,8,†]

[1]Alevel Center, Jinling High School, Nanjing, 210005, China
[2]Business college, Hunan Agricultural University, Changsha, 410128, China
[3]Mount Holyoke College, Hadley, 01075, The USA
[4]International Academy, Shenzhen Foreign Language School, Shenzhen, 518083, China

[5]1013469895@qq.com
[6]jiayi.zhu121@gmail.com
[7]ding224j@mtholyoke.edu
[8]liziyan20220913@163.com
[9]Corresponding author email: 1013469895@qq.com
[†]All the authors contributed equally to this work and should be considered as co-first author.

Abstract. In the past, analysts evaluated whether to offer loans to particular applicants using rule-based approaches. However, due to the sudden rise in applicants and a labor shortage, financial institutions have created quantitative methods of decision-making. Credit scoring models are constructed. In this essay, random forest model, support vector machine regression model and Probit model are performed and compared according to the dataset from a major U.S. credit cards company. The result demonstrates that while machine learning techniques can improve the efficiency and accuracy of credit risk assessment, it does face some problems and limitations. Random forest model is capable of handling high-dimensional data and is not complicated to run. However, database with fewer features or samples will have lower classification accuracy. Support vector machine regression model has high accuracy and prevents overfitting to some degree. It is sensitive to the choice of kernel parameters and regularization term. By testing how important Mill Ratio is, the Probit model produces more accurate results. However, the model is more complex than the other two. In future research, we propose to enhance and extend our work by using more artificial intelligence algorithms and evaluation metrics.

Keywords: credit risk model, artificial intelligence, random forest, support vector machine, Probit model

1. Introduction

Research on artificial intelligence (AI) in analyzing bad debt rates has been widely studied. In the past, analysts used rule-based approaches (such as the 5C categorization approach) to assess whether to grant loans to specific applicants. However, with the rapid increase in the number of applicants and a shortage of manpower, financial institutions have developed quantitative methods to make decisions. For example, they have developed credit scoring models to classify applicants as either accepted or rejected based on their characteristics. The goal of credit scoring models is to allocate credit applicants into the "good credit" group, which is likely to fulfill their financial obligations, or the "bad credit" group, whose applications will be rejected due to a higher likelihood of defaulting on their financial obligations [1]. The use of AI in solving classification problems to support credit decisions is gradually playing an important role in business-related decisions [2].

However, in practice, AI technologies in credit risk assessment also face difficulties and challenges, such as data quality, data security, model interpretability, model stability, and model bias. The success rate of AI is influenced by specific factors. If AI technology makes errors or mistakes in credit risk assessment, it could lead to significant economic losses and legal liabilities for banks and financial institutions, and even impact the stability and trustworthiness of the entire financial system. This paper aims to compare and demonstrate the issue of overfitting in AI when analyzing bad debt rates by building decision tree model, support vector machine (SVM) model, and Probit model. We will utilize these models to explore the impact of overfitting issues on bad debt rate predictions and propose corresponding solutions and improvement methods. A study on methods for predicting financial crises found that these methods can be categorized into bottom-up approaches, aggregate approaches, and macroeconomic approaches [3]. In the research of forecasting models for the Indian banking industry crisis, the author utilized an ordered probit model and incorporated macroeconomic indicators as predictive factors.

The organization of this paper is as follows. In the second part, we summarize typical research on credit risk models. The third part describes the methods, experimental design, and data collection and analysis process used in our study. The fourth part shows data visualization of the dataset and variables we choose for the models. The fifth part gives people a clear view on limitations on models we use. The sixth section shows up conclusions – results we gain and learn from out models. Last but not list is the Bibliography section.

2. Literature review

In machine learning, the accuracy of different models in predicting default rates is an important issue of concern for researchers worldwide. This article reviews several relevant studies with the aim of comparing the accuracy of different credit risk models.

One study on credit scoring ensemble models was conducted by Ghodselahi et al [4]. They employed ensemble learning techniques, using 10 classifier agents as members of the ensemble model, and compared the classification accuracy of SVM, neural networks, and decision trees as base classifiers. The experimental results showed that the ensemble model had advantages in classification accuracy and performance compared to other credit scoring methods. However, it should be noted that the ensemble model may face overfitting issues, where the performance of the entire model may decline if the member classifiers are too complex or overfit. Another study by Niklis et al. [5] utilized non-parametric machine learning techniques based on the SVM framework for building a credit rating system model. The research results indicated that the additive SVM model yielded the best results among the machine learning techniques used. This further demonstrates that using SVM methods can effectively enhance the performance of credit scoring models. Golbayani et al. [6] conducted an investigation and comparative analysis of literature results pertaining to the application of machine learning techniques in predicting credit ratings. They applied four machine learning techniques-bagged decision trees, random forests, SVM, and multilayer perceptron - to the same dataset and compared the experimental results. The research findings showed that models based on decision trees exhibited superior performance. This further validates the effectiveness of decision trees in credit scoring. A recent study by Anand et al. [7] utilized loan data from multiple internet sources as well as a dataset of loan applications from applicants to propose various techniques for computing important indicators of credit scores, combining random forests, SVM, and other ensemble methods. Through measures such as accuracy, F1 score, ROC analysis area, and feature importance, the research results demonstrated the effectiveness of ensemble methods such as random forests and SVM in credit rating. Although machine learning algorithms might be questioned by the outer world on its accuracy ability when working for data analysis in banking system or other financial services, some researchers still believe that machine learning is more efficient in analyzing a large number of loan applications to derive accurate results on the non-approval of credit for vulnerable consumers since AI can be less time consuming, money wasting, and even being able to be applied in complicated real world conditions. For example, Crook et al. [8] found that PD models use classification models based on statistical or machine learning methods to predict borrower default. The advantage of statistical methods lies in their ability to quantitatively display the impact of various factors on borrower default. Overall, machine learning performs better than statistical learning methods in terms of predicting default rates. Machine learning can quickly present more possibilities and incorporate or exclude more factors.

3. Methodology

This study uses a combination of Random Forest to build and evaluate an AI-based credit risk assessment model, and use k-fold Cross Validation to improve accuracy.

Random Forest (RF) refers to the establishment of a forest by random sampling where the forest refers to many independent decision trees.

The basic principle of random forest is as follows:

First: select K features from the dataset and clean the data.

Second: build N decision trees based on the chosen features as a random forest classifier Third: train the classifier and predict the outcome.

Fourth: use K-fold cross-validation to score this classifier.

Fifth: repeatedly build m decision trees based on these K features.

Finally: compare the classifiers and get the best one.

K-fold cross-validation is a method used to evaluate model performance. Its basic idea is to divide the original data set into K parts, select one part as the testing set each time and the remaining K-1 parts are the training sets. For the training set, repeat K times to get the test results, then take the average as final evaluation result.

SVM is a classification algorithm whose purpose is to find an optimal hyperplane in a given data set, so that the data points of different categories are separated as much as possible, and the data points closest to the hyperplane (called support vectors) has the largest distance to the hyperplane. This improves classification accuracy and generalization.

The basic steps of the SVM algorithm are as follows:

- First, assume that the data set is linearly separable, that is, there exists a linear equation.

$$\boldsymbol{w} \cdot x + b = 0$$

It can divide the data points into two categories, where $\boldsymbol{w}$ is the normal vector of the hyperplane, b is the intercept of the hyperplane, and x is the eigenvector of the data points.

- Then, define the geometric interval of the hyperplane about the data point (x_i, y_i) as $\gamma i = y_i\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{xi} + \frac{b}{\|\boldsymbol{w}\|}\right)$where y_i is the class label of the data point and takes the value +1 or -1. The geometric interval represents the signed distance of the data point to the hyperplane, with the sign consistent with the class label.
- Next, solve the problem of maximizing the minimum$\gamma = \mathrm{mi\,n\,i} = 1,2 \ldots, \mathrm{N}\gamma i$ of the geometric spacing of all data points, which is equivalent to solving the problem of maximizing $\frac{1}{\|\boldsymbol{w}\|}$is also equivalent to solving the problem of minimizing$\frac{1}{2} \| \boldsymbol{w} \|^2$ while satisfying the constraints $y_i(\boldsymbol{w} \cdot \boldsymbol{xi} + b) \geq 1, i = 1,2, \ldots, N$.
- Finally, using the Lagrange multiplier method and KKT conditions, the original problem is transformed into a dual problem, and the optimal solution is obtained by solving the dual problem. The dual problem is a quadratic programming problem about the Lagrange multiplier αi, and its objective function is below.

$$\boldsymbol{w} \cdot x + b = 0$$

The optimal Lagrangian multipliers and hyperplane parameters can be obtained by solving the dual problem, and it can be found that only some data points correspond to Lagrangian multipliers that are not zero, and these data points are support vectors.

Probit Model

Probit model is used for binary dependent variables. In the given dataset, scstars_bk_v1, I choose BADONUS and BADOFFUS as my dependent variables, respectively. Probit model helps calculating out the linear prediction with a standard normal mean 1. We are supposed to know what the area is right and left to the number “0”. There is a pattern that when fitted value increases, the distribution shifts right, the probability of the observing value would increase if we increase the fitted value.

The basic formula for Probit model is y=B0+B1X1+B2X2+…+BkXk+U, the formula represents latent variable contains the linear prediction plus an error term that has variance of 1 and mean of 0.

Mill Ratio (MR) is seen as an extra condition in a scenario. Mill ratio is calculated by normal PDF divided by normal CDF. Then comparing the probability with Mill ratio with the one without Mill Ratio.

4. Data visualization

4.1. Data processing

First, Read the data in R.

Second, Choose the columns that we need in our models.

Third, Delete rows with MISSING VALUES.

Independent variables I choose areChoose column [1,2,8,9,44] dependent variables (BADONUS & BADOFFUS) & independent variables (AgeAvgTrd + AgeYoungTrd + BRAgeAVG).

AgeAvgTrd is Average Age of all trades in months.

AgeYoungTrd is Age of youngest trade.

BRAgeAVG is Average Age of Bank Revolving Trade.

4.2. Data visualization

Fig. 1-5 show the result of Data visualization.

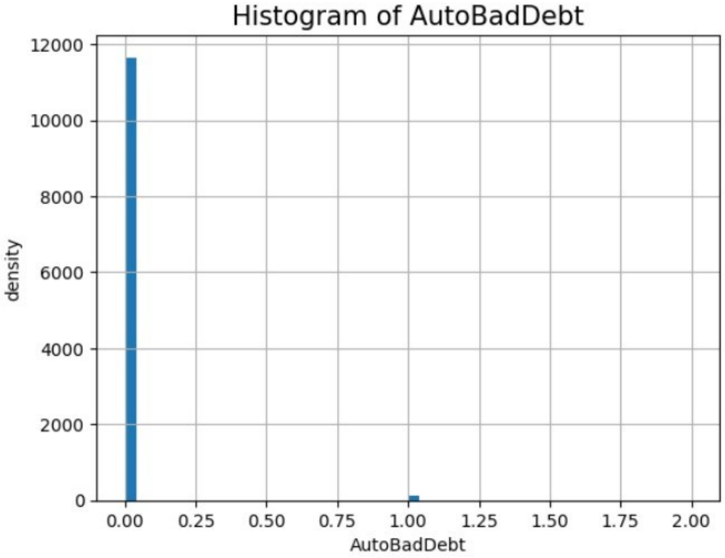


Figure 1. Histogram of AutoBadDebt.

The histogram of AutoBadDebt demonstrates that the data within AutoBadDebt is discrete. Obviously, the quantity of 0 far exceeds the quantity of 1. And the number of 2s is so small that it is not even clearly represented on the graph.

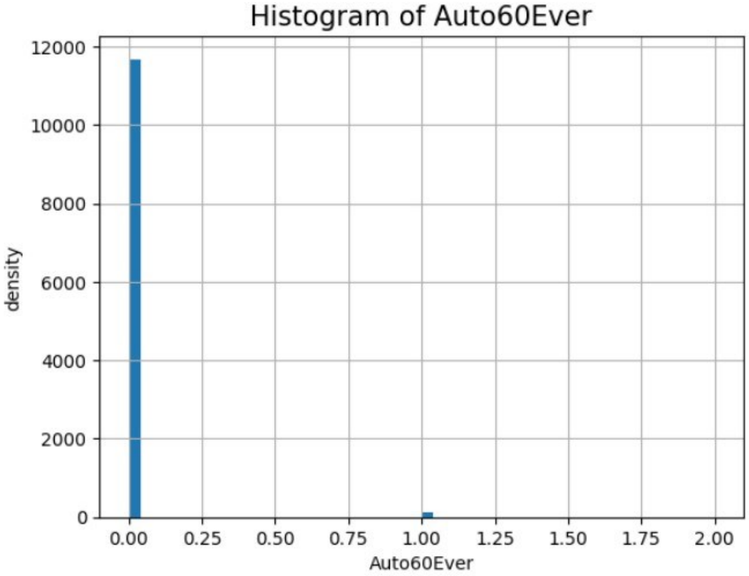


Figure 2. Histogram of Auto60Ever.

The histogram of Auto60Ever demonstrates that the data within Auto60Ever is discrete. Obviously, the quantity of 0 far exceeds the quantity of 1. And the number of 2s is so small that it is not even clearly represented on the graph.

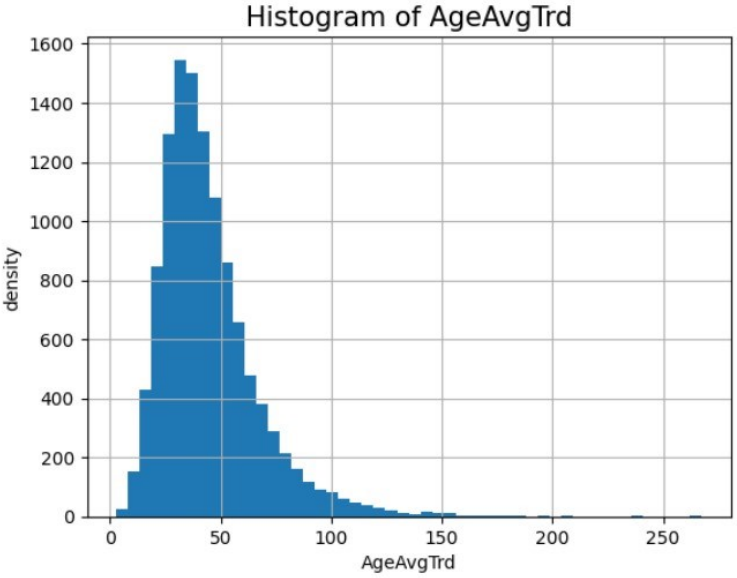
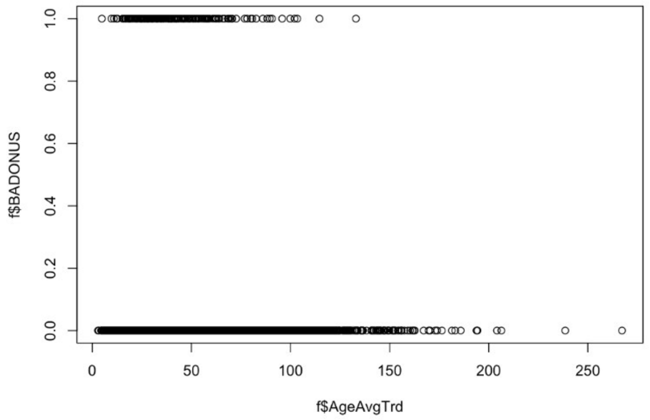


Figure 3. Histogram of AgeAvgTed.

The histogram of AgeAvgTed demonstrates that the data within AgeAvgTed is continuous and has positive skewness.

□ AgeAvgTrd & BADONUS

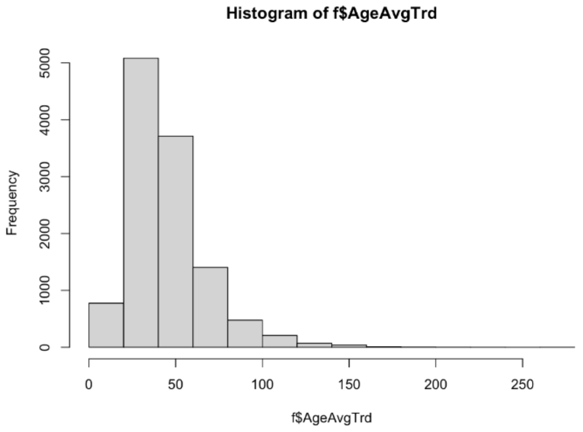


(source: AgeAvgTrd & BADONUS plot in R)

Figure 4. AgeAvgTrd and BADONUS plot.

The graph shows a bigger amount that AgeAvgTrd causes BADONUS than it does not have any impact on BADONUS.

□ Histogram of f$AgeAvgTrd



(source: Histogram of f $ AgeAvgTrd in R)

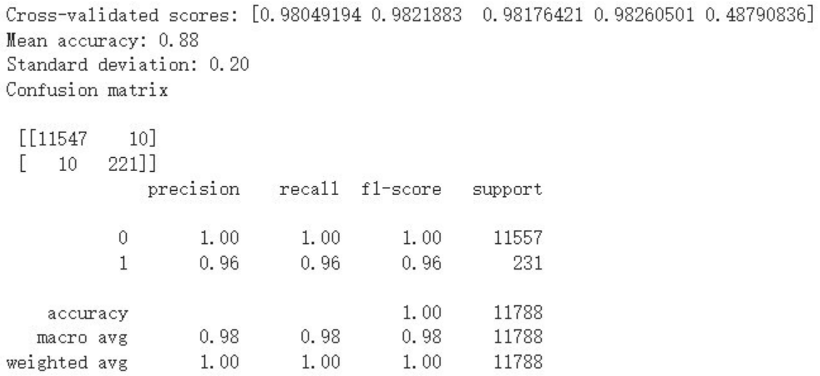
Figure 5. Histogram of AgeAvgTrd.

The graph above shows a positively skewness graph of AgeAvgTrd

4.3. Model building

Two models that are based on the classification theory are taken into consideration because the cleaned dataset contains over 10,000 samples and several attributes.

The study first uses random forest to build and evaluate an AI-based credit risk assessment model, and use the Cross Validation to improve accuracy.

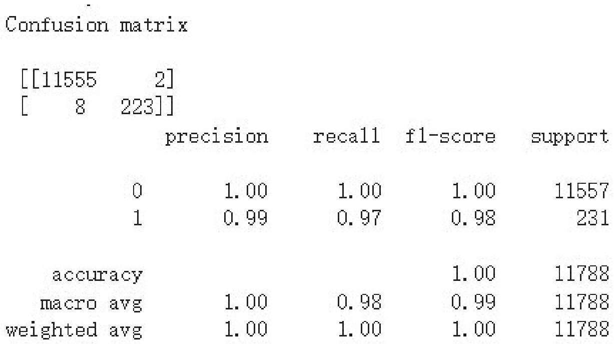


```
Cross-validated scores: [0.98049194 0.9821883  0.98176421 0.98260501 0.48790836]
Mean accuracy: 0.88
Standard deviation: 0.20
Confusion matrix

 [[11547    10]
 [   10   221]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     11557
           1       0.96      0.96      0.96       231

    accuracy                           1.00     11788
   macro avg       0.98      0.98      0.98     11788
weighted avg       1.00      1.00      1.00     11788
```

Figure 6. The confusion matrix of using 5 decision trees.

At first, 5 decision trees are built as a classifier and the classified results are shown above (Fig.6). 20 pieces of data are misclassified. Additionally, a relatively small number of 1s in BADONUS results in a worse classification accuracy than a big number of 0s.

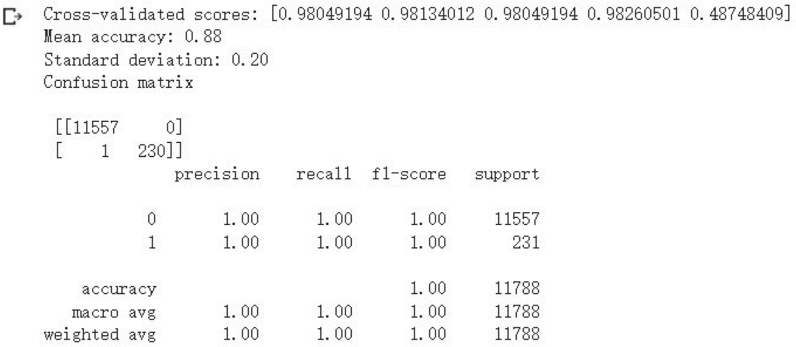
[The data within BADONUS only contains 1s and 0s while the quantity of 0 far exceeds the quantity of 1. (The number of people that have bad debts is not the majority)]



```
Confusion matrix

 [[11555     2]
 [    8   223]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     11557
           1       0.99      0.97      0.98       231

    accuracy                           1.00     11788
   macro avg       1.00      0.98      0.99     11788
weighted avg       1.00      1.00      1.00     11788
```

Figure 7. The confusion matrix of using 10 decision trees.

10 decision trees are built and the accuracy has improved. Only 10 pieces of data are incorrectly predicted (Fig.7).



```
Cross-validated scores: [0.98049194 0.98134012 0.98049194 0.98260501 0.48748409]
Mean accuracy: 0.88
Standard deviation: 0.20
Confusion matrix

 [[11557     0]
 [    1   230]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     11557
           1       1.00      1.00      1.00       231

    accuracy                           1.00     11788
   macro avg       1.00      1.00      1.00     11788
weighted avg       1.00      1.00      1.00     11788
```

Figure 8. The confusion matrix of using 20 decision trees.

20 decision trees are built and the results are shown above (Fig.8). And there is almost no misclassified data.

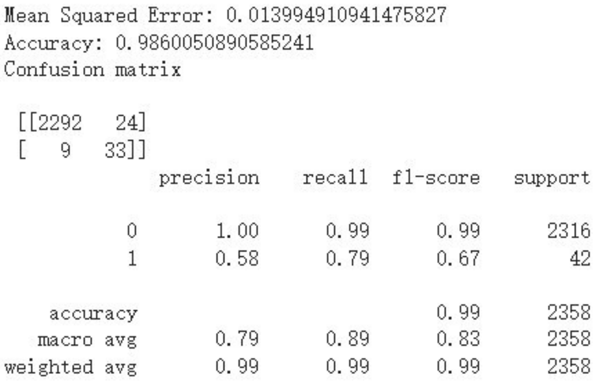
Through the classification reports of 5, 10, 20 decision trees respectively, it is found that the accuracy of random forest depends on the number of decision trees that built.

SVM Model Building:

Another regression model is based on the mathematical theory. SVM is a classification algorithm which purpose is to find an optimal hyperplane in a given dataset, so that the data points of different categories are separated as much as possible.

The basic steps of analyzing the data by using SVM algorithm is as shown:

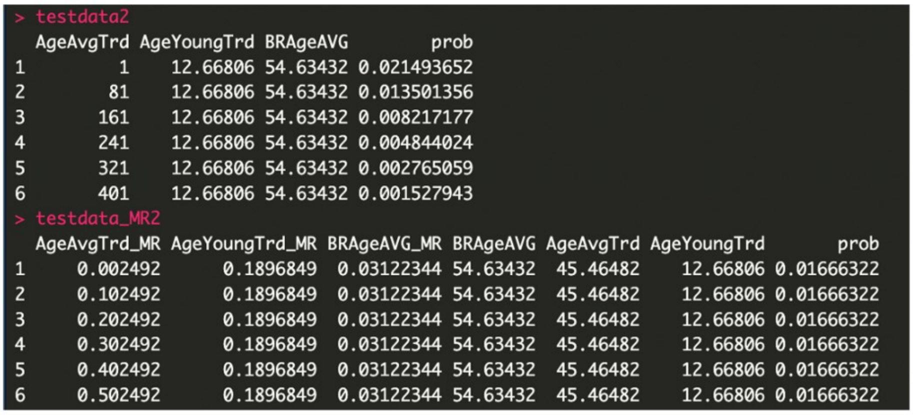
The data in the cleaned dataset is first normalized between 1 and 0. In order to prevent overfitting, it is divided into a training set and a testing set. The third step is the construction of an SVM regression model using the SVM algorithm that was introduced in the methodology section. Finally, I train the regression model and make an outcome prediction.



```
Mean Squared Error: 0.013994910941475827
Accuracy: 0.9860050890585241
Confusion matrix

 [[2292   24]
 [   9   33]]
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      2316
           1       0.58      0.79      0.67        42

    accuracy                           0.99      2358
   macro avg       0.79      0.89      0.83      2358
weighted avg       0.99      0.99      0.99      2358
```

Figure 9. The confusion matrix of SVM regression model.

The confusion matrix of SVM regression model shows that it isn't suitable for large number of data samples (Fig.9).

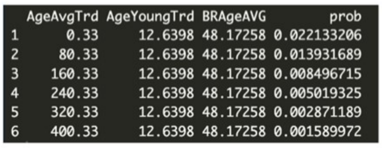


```
> testdata2
  AgeAvgTrd AgeYoungTrd BRAgeAVG        prob
1         1    12.66806 54.63432 0.021493652
2        81    12.66806 54.63432 0.013501356
3       161    12.66806 54.63432 0.008217177
4       241    12.66806 54.63432 0.004844024
5       321    12.66806 54.63432 0.002765059
6       401    12.66806 54.63432 0.001527943
> testdata_MR2
  AgeAvgTrd_MR AgeYoungTrd_MR BRAgeAVG_MR BRAgeAVG AgeAvgTrd AgeYoungTrd       prob
1     0.002492      0.1896849  0.03122344 54.63432  45.46482    12.66806 0.01666322
2     0.102492      0.1896849  0.03122344 54.63432  45.46482    12.66806 0.01666322
3     0.202492      0.1896849  0.03122344 54.63432  45.46482    12.66806 0.01666322
4     0.302492      0.1896849  0.03122344 54.63432  45.46482    12.66806 0.01666322
5     0.402492      0.1896849  0.03122344 54.63432  45.46482    12.66806 0.01666322
6     0.502492      0.1896849  0.03122344 54.63432  45.46482    12.66806 0.01666322
```

(source: probability of testdata with MR and without MR)

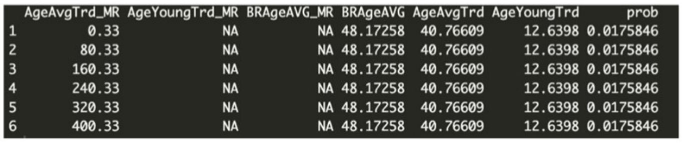
Figure 10. Probability of testdata with MR and without MR.

When setting BADONUS as the dependent variable, AgeAvgTRd, AgeYoungTrd, BRAgeAVG as the three independent variables. Dividing the three independent variables into 6 different groups by their number range and calculating their probabilities. For probabilities with Mill Ratio, AgeAvgTrd seems to have a bigger correlation since the probability is bigger compared to the probabilities in the 6 categories without MR (Fig.10).



```
  AgeAvgTrd AgeYoungTrd BRAgeAVG        prob
1      0.33     12.6398 48.17258 0.022133206
2     80.33     12.6398 48.17258 0.013931689
3    160.33     12.6398 48.17258 0.008496715
4    240.33     12.6398 48.17258 0.005019325
5    320.33     12.6398 48.17258 0.002871189
6    400.33     12.6398 48.17258 0.001589972
```

(source: probibility without MR in 6 degrees)



```
  AgeAvgTrd_MR AgeYoungTrd_MR BRAgeAVG_MR BRAgeAVG AgeAvgTrd AgeYoungTrd      prob
1         0.33             NA          NA 48.17258  40.76609     12.6398 0.0175846
2        80.33             NA          NA 48.17258  40.76609     12.6398 0.0175846
3       160.33             NA          NA 48.17258  40.76609     12.6398 0.0175846
4       240.33             NA          NA 48.17258  40.76609     12.6398 0.0175846
5       320.33             NA          NA 48.17258  40.76609     12.6398 0.0175846
6       400.33             NA          NA 48.17258  40.76609     12.6398 0.0175846
```

(source: probibility MR in 6 degrees)

Figure 11. Probability of testdata in 6 degrees.

Even changing the Y value into BADOFFUS, Mill Ratio (MR) in BADOFFUS also shows a closer relationship of AgeAvgTrd and BADODOFFUS (Fig.11).

In a word, we could conclude that Mill Ratio is important. Mill Ratio can be seen as different scenarios and conditions in real world. When we are evaluating something, we should always consider the Mill Ratio. Machines are helping us figuring out results in a more efficient and money saving way. Imagining implementing different conditions we expect into real life, the result can be wrong and at the same time countless money and time are wasted.

KS-Test is the one we used for checking whether the model is appropriate for the dataset or not.

We use AgeAvgTrd & BRAgeAVG and find out that their test statistic is 0.1204 and their corresponding p-value is 2.2e-16 which is smaller than 0.05, so null hypothesis is rejected. So, the two sample datasets do not come from the same distribution, then the model works.

5. Discussion and Limitations

5.1. Advantages and disadvantages for Random Forest

Several advantages of random forest are found through the process. First and foremost, it is capable of handling high-dimensional data. With random forest, the process is not severely influenced if a new piece of data is added to the dataset or if one piece of data is incorrect. (It will only affect one decision tree and is difficult to affect all decision trees.) Moreover, it is not complicated to run and has high training speed.

However, database with fewer features or samples will have lower classification accuracy. Additionally, the internal operation of the model is unable to control by the experimenters. When there are many decision trees, overfitting can become a problem, although k fold cross validation can help. It is also possible to cut some brunches of the decision trees.

5.2. Advantages and disadvantages for SVM

SVM has several advantages, such as it has high accuracy and performs well in high-dimensional spaces. Second, it has regularization capabilities that prevent overfitting and improve generalization. Third, it can handle both linear and non-linear data using different kernels. Fourth, it is robust to outliers and noise in the data.

However, SVM also has some limitations, such as it is sensitive to the choice of kernel parameters and regularization term, which require tuning and cross-validation. And it may computationally expensive and time-consuming for large datasets, especially when using non-linear kernels. Additionally, it does not provide probability estimates for the predictions, unlike some other algorithms such as logistic regression or naive Bayes. Or it is difficult to interpret and explain the results, especially when using complex kernels.

5.3. Advantages and disadvantages for Probit Model

The advantages for Probit model are: First, getting more accurate results by testing how important MR; Second, avoiding ignoring different consequences & conditions

The disadvantages are: First, the model complexity compared to Random Forest is high since multiple groups of Y value and X values are supposed to be compared and divided into 6 different categories; Second, it is easy making mistakes on not setting up the controlled group correctly; Third, before finding out the correct method, a bunch of time are wasted on Binning. However, the dataset does not fit binning method since there are too many “zeros” and very little “ones”.

6. Conclusion

As the cleaned dataset contains over 10,000 samples and several attributes, two models that are based on the classification algorithm are taken into consideration. The random forest is served as the first model. Different numbers of decision trees are constructed as classifiers and are trained to forecast the classification results. Another regression model built on the SVM algorithm aims to separate data into two categories in a hyperplane as far as possible. The prediction reports are displayed as confusion matrices.

Probit model shows clear probability comparison between the two different groups – the one with Mill Ratio and the one without. The result is that Mill Ratio helps considering extra conditions in real life and increase the probability between independent variables and dependent variable. KS-test gives us a value smaller than 0.05 which further confirms that Probit Model works for the dataset. In a word, machine learning helps figuring real world condition as well since it gives people hints on something they could have never expected of. Making a control group in machine learning gives people a brighter view on what decisions they should make which saves both time and money.

In this paper, we explore the application and limitations of AI in bank credit risk assessment. We find that AI helps us reduce financial risks by giving clear data preview and saving us time and money at the same time. We use random forest, SVM and Probit model to build credit risk classification models, to prove the imperfection of AI credit classification.

Our research provides some valuable insights for bank credit risk assessment, but also has some shortcomings and limitations. In future research, we suggest to improve and expand from the following aspects:

First, we can try to use more data and features to train and test our models, to improve the accuracy and generalization ability of the models. We can collect more credit risk related information from different data sources and channels, such as customer's personal information, financial situation, consumption behavior, social network, etc., to increase the input dimension and complexity of the models.

Second, we can try to use more AI algorithms and techniques to build and optimize our models, to improve the efficiency and flexibility of the models. We can explore some emerging AI fields and methods, such as deep learning, reinforcement learning, neural network, etc., to enhance the learning ability and adaptability of the models.

Third, we can try to use more evaluation indicators and methods to evaluate and compare our models, to improve the reliability and interpretability of the models. We can measure the performance of the models from different angles and levels, such as accuracy, recall, F1 value, AUC value, ROC curve, etc., to comprehensively reflect the advantages and disadvantages of the models.

Finally, we can try to combine our models with other fields and scenarios, to improve the practicality and innovation of the models. We can apply our models to other types or sizes of banks or financial institutions, or extend our models to other fields or industries, such as insurance, loans, investment, etc., to explore the wider and deeper application of AI in credit risk assessment.

Through these aspects of improvement and expansion, we hope to provide more effective and advanced AI solutions for bank credit risk assessment, and also provide more inspiration and reference for AI applications in other fields and scenarios.

References

[1] Lee TS, Chiu CC, Chou YC and Lu CJ. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis* 2006, 50: 1113–1130.

[2] Thomas LC. A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting* 2000, 16: 149–172.

[3] Singh TR. An ordered probit model of an early warning system for predicting financial crisis in India. *IFC Bulletin* 2011, 34: 185-201.

[4] Ghodselahi A and Amirmadhi A. Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modeling and Optimization* 2011, 1: 243-249.

[5] Niklis D, Doumpos M and Zopounidis C. Combining market and accounting-based models for credit scoring using a classification scheme based on support vector machines. *Applied Mathematics and Computation* 2014, 234: 69-81.

[6] Golbayani P, Florescu I and Chatterjee R. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance* 2020, 54: 101251.

[7] Anand M, Velu A and Whig P. Prediction of loan behaviour with machine learning models for secure banking. *Journal of Computer Science and Engineering* 2022, 3: 1-13.

[8] Crook JN, Edelman DB and Thomas LC. Recent developments in consumer credit risk assessment. *European Journal of Operational Research* 2007, 183: 1447-1465.