

# Predicting of Credit Default by SVM and Decision Tree Model Based on Credit Card Data

Jiaqi Fan

Fudan University, Shanghai, China

20307110351@fudan.edu.cn

**Abstract.** With the global financial crisis and increased credit risk, default forecasting is playing an increasingly important role in every sector of the economy. Currently, there are linear models and machine learning models for predicting credit defaults. In recent years, big data risk control models are superior to traditional bank models in predicting default rates, and can also conduct business quickly and on a large scale. This paper compares the SVM and the decision tree model in the machine learning model based on the credit card loan data set, and finally evaluates the prediction effect between the two models. According to the study, the decision tree model outperforms the SVM in terms of prediction accuracy. The use of big data to conduct machine learning to predict credit conditions enables financial institutions to serve small, medium and micro enterprises that were difficult to cover by traditional finance on a large scale in the past. It is a world-class innovation in finance.

**Keywords:** Credit Default, forecasting, SVM, Decision Tree.

## 1. Introduction

As an objective measure of market default and credit risk, the default rate is of great significance to all participants in the bond market. In mature foreign bond markets, default rates are widely used [1-2]. With the continuous maturity of China's bond market, the default rate in the domestic market has also received more and more attention. Credit rating agencies should develop a long-term rating quality verification mechanism with default rates as the core, according to a recent file named "Notice of the PBC, the NDRC, the MFCBIRC, and CSRC on Promoting the Healthy Development of the Bond Market Credit Rating Industry." As the core indicator of risk monitoring and rating quality inspection, default rate is widely used in level inspection and risk measurement [3].

The default rate refers to the historical default frequency obtained from the statistics of the actual bond defaults of the rated entity. Default rate is not a single concept, but a collection of default-related concepts and calculation indicators. According to different application scenarios, the use and interpretation of default rates in practice are also diversified, mainly reflected in the differences in default rate categories and calculation calibers.

With the global financial crisis and increased credit risk, corporate default forecasting is playing an increasingly important role in every sector of the economy. In order to analyze credit risk, digital financial institutions can potentially utilize machine learning techniques to replace mortgage assets with data such as real-time transactions and behavioral traits [4-5]. Using machine learning methods to predict default rates is conducive to banks' reasonable credit assessment.

In this paper, the machine learning model is innovatively applied, and the advantages of big data are brought into. Machines can self-learn a large amount of financial and non-financial data to mine massive, multi-dimensional and dynamic data information to improve the prediction accuracy. The article found that the decision tree model performs better in prediction accuracy than SVM, and has lower RMSE and MSE.

The article is organized in the following. The data make up the Section 2. The research methodology is the third section. The fourth section is an examination of the research findings, and the conclusion comes last.

## 2. Methodology

This paper mainly compares the SVM and the decision tree model, and explores the effect of different models on the default rate prediction. The following two models will be briefly introduced.

### 2.1 SVM

A generalized linear classifier known as a support vector machine (SVM) conducts binary classification of data using the supervised learning approach. The greatest margin hyperplane that resolves the planning concerns in the learning sample serves as its decision boundary. Support vector machines offer a more efficient and effective method of learning complex nonlinear equations when compared to logistic regression and neural networks [6-7]. It is crucial to find the ideal classification hyper-plane of the two categories of samples in the original space when they are linearly separable. When it is initially linearly inseparable, the addition of slack variables and the use of nonlinear mapping to transfer samples from the low-dimensional input space to the high-dimensional space make it separable linearly, allowing the best classification hyperplane to be located in the feature space [8]. The equation of SVM is as follows.

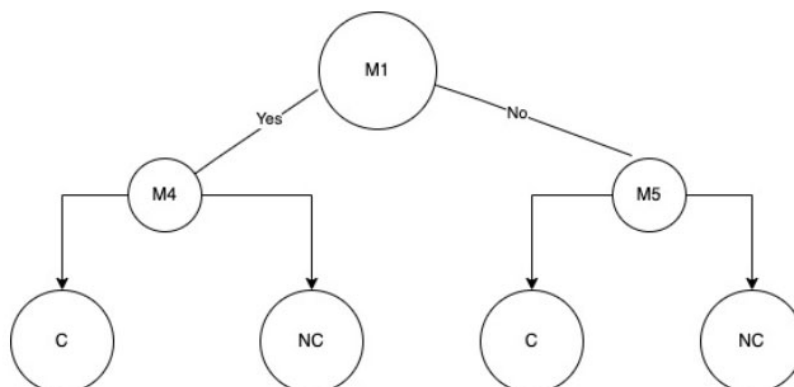
$$\min \frac{1}{2} |w|^2 \quad (1)$$

$$s.t. y_i (w^T X_i + b) \geq 1 \quad (2)$$

The following are some benefits of SVM in forecasting. (1) Small sample machine learning issues can be resolved using the support vector machine approach, which also makes common classification and regression issues easier to understand. (2) When mapping to a high-dimensional space, the employment of the kernel function approach does not increase the computational complexity because it overcomes the drawbacks of dimensionality and nonlinear reparability. In other words, the computational complexity of the support vector algorithm depends on the number of support vectors rather than the size of the sample space because the final decision function is only controlled by a small number of support vectors.

### 2.2 decision tree model

A decision tree is both a descriptive and predictive model that categorizes unidentified events and evaluates which cases are indistinguishable from various groups. On variable that affects learning, decision tree classifiers are produced. The fundamental of decision tree learning may be generating a set of classification rules from the training set or estimating a conditional probability model from the training set [9–10]. Questions and responses judgment and decision trees both rely on the same principles. The issue is answered once it is verified whether or not to be accurate based on a set of data. As a consequence, the decision tree classifier's applicability is improved. This objective is represented by a loss function in decision tree learning, which is often a regularized maximum likelihood function. Decision tree learning uses the loss function's minimization as the goal function. The structure is as follows (Fig 1).

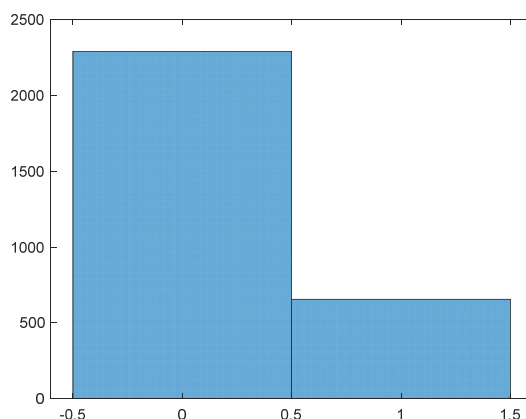


**Fig. 1** Decision tree model

### 3. Experiments and Results

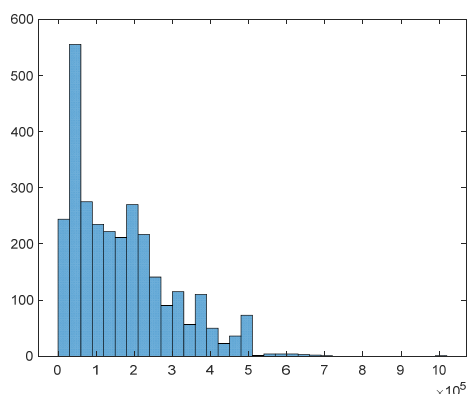
#### 3.1 Data

This article explores the impact of factors such as age, gender, education, and loan amount on the default rate. The article uses the UCI Credit Card Fraud dataset to describe the statistics of each variable. Finally, the machine learning method is used to consider the impact of each variable on the default prediction.

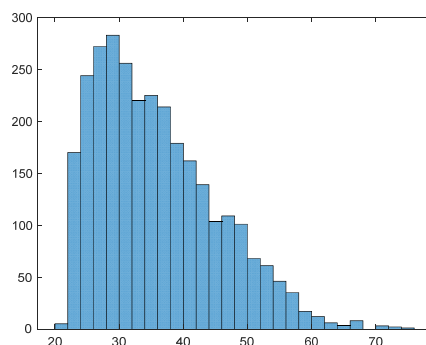


**Fig. 2** Default situation

According to Fig 2, about three-quarters made their repayments on time, and about one-quarter were late on their payments and had credit defaults. Based on Fig 3, The majority of borrowings are under \$500,000, and a few are over \$500,000. The largest number of borrowers are those with a loan of \$20,000.

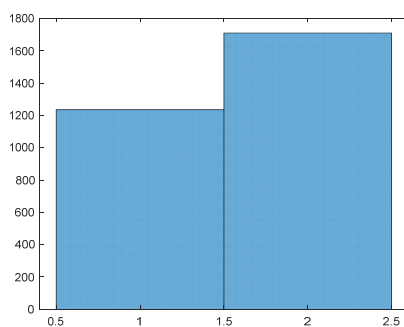


**Fig. 3** amount of loan.



**Fig. 4** Age distribution

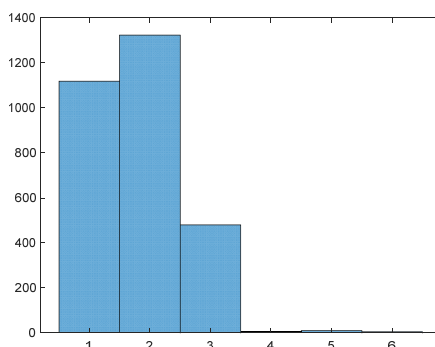
Figure 4 shows the age distribution of these nearly 3,000 samples, of which the 30-year-old is the largest. The second is the 25-30-year-old group. The overall distribution obeys a normal distribution. The oldest is 80 years old, and they are also using credit cards to make a loan.



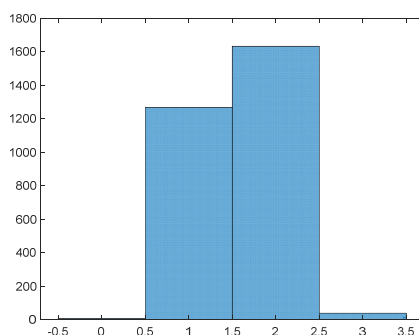
**Fig. 5** Gender distribution

Figure 5 summarizes gender. Among them, 1 represents women and 2 represents men. The article found that the majority of the population in the sample were men, and women used credit card loans less.

Figure 6 represents the educational level of different people's distribution of the dataset. The higher the number, the higher the educational level. 1, 2, 3 denote elementary school, junior high, and high school, respectively. 4 and above means undergraduate or above. This paper finds that most of the people belong to junior high school. Figure 7 shows the marital status of the data sample. Among them, 1 represents married, and 2 represents married. It can be found that most borrowers belong to the married group.



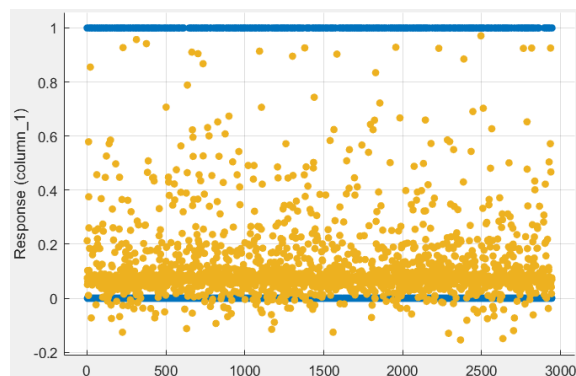
**Fig. 6** Education level by different number



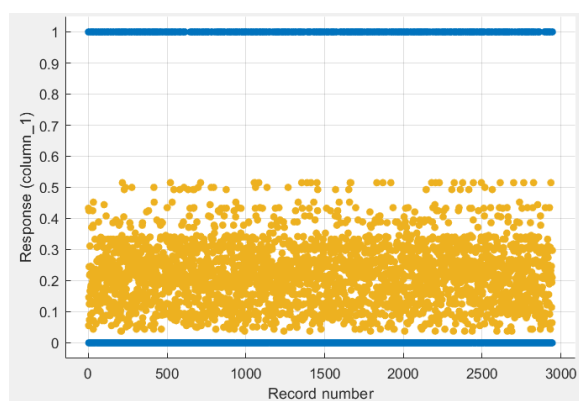
**Fig. 7** Marriage condition

### 3.2 Results analysis

This paper considers SVM and Decision tree model to predict the default condition of the sample based on a series of features like age, gender, education, and loan amount. Figure 8 and 9 summarize the final results and prediction results of SVM and Decision tree model. The yellow point represents the prediction result. The evaluation result of different models is shown in Table 1. RMSE, MAE and MSE are commonly used to evaluate the forecasting efficiency of model. The number of the indicator is small, representing that the performance of the model is better.



**Fig. 8** Results of SVM



**Fig. 9** Results of decision tree

**Table 1.** comparison of SVM and Decision tree.

	SVM	Decision tree
RMSE	0.4464	0.4222
MSE	0.1992	0.1782
MAE	0.2808	0.3416

RMSE, MAE and MSE are commonly used to evaluate the forecasting efficiency of model. The number of the indicator is small, representing that the performance of the model is better. Based on the result of Table 1. The RMSE and MSE of SVM are 0.4464 and 0.1992. Besides, Decision tree model are 0.4222 and 0.1782, respectively, which are lower than SVM. Thus, the Decision tree model is better in forecasting credit default rate than SVM. Based on the literature, MAE is not so good indicator compared to RMSE and MSE. Thus, that SVM's MAE is lower than decision may not show its robustness [10]. We conclude that Decision tree model of machine learning is better in predicting credit default rate.

## 4. Conclusion

The measurement of the probability of default is vital to and necessary to commercial banks' management of credit risk. The first criteria for credit risk management is something like this. The significance of credit rating is based on the assessment of the borrower's risk of defaulting as a primary method for assessing credit risk. Only after estimating the borrower's probability of default properly can the bank estimate the probable loss and generate a precise assessment of the customer's credit standing, confirming the effectiveness and scientific nature of commercial banks' credit risk management. Predicting default rates is thus especially essential.

Based on the machine learning model (SVM and decision tree model), this paper uses the UCI Credit Card Fraud dataset to predict the default rate of users, considering age, gender, education, marriage status and loan amount as feature variables. The article found that the decision tree model performs better in prediction accuracy than the SVM, and has lower RMSE and MSE.

There are still some deficiencies in this paper. Only 3000 samples are considered in the sample selection, and the number is insufficient, which is not conducive to the training of machine learning. In the future, this paper will consider more samples for prediction.

## References

- [1] Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016, August). A review of machine learning techniques using decision tree and support vector machine. In 2016 international conference on computing communication control and automation (ICCUBE) (pp. 1-7). IEEE.
- [2] Nie, F., Zhu, W., & Li, X. (2020). Decision Tree SVM: An extension of linear SVM for non-linear classification. *Neurocomputing*, 401, 153-159.
- [3] Broadstock, D. C., Chan, K., Cheng, L. T., & Wang, X. (2021). The role of ESG performance during times of financial crisis: Evidence from COVID-19 in China. *Finance research letters*, 38, 101716.
- [4] Egorova, A. A., Grishunin, S. V., & Karminsky, A. M. (2022). The Impact of ESG factors on the performance of Information Technology Companies. *Procedia Computer Science*, 199, 339-345.
- [5] Folger-Laronde, Z., Pashang, S., Feor, L., & ElAlfy, A. (2022). ESG ratings and financial performance of exchange-traded funds during the COVID-19 pandemic. *Journal of Sustainable Finance & Investment*, 12(2), 490-496.
- [6] Ademi, B., & Klungseth, N. J. (2022). Does it pay to deliver superior ESG performance? Evidence from US S&P 500 companies. *Journal of Global Responsibility*
- [7] Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2016). Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, 12(3), 1005-1016.
- [8] Takahashi, F., & Abe, S. (2002, November). Decision-tree-based multiclass support vector machines. In *Proceedings of the 9th International Conference on Neural Information Processing*, 2002. ICONIP'02. (Vol. 3, pp. 1418-1422). IEEE.
- [9] Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*, 115, 29-37.
- [10] Sun, L., Fu, S., & Wang, F. (2019). Decision tree SVM model with Fisher feature selection for speech emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019(1), 1-14.