Research on the Application of Machine Learning Algorithms in Credit Risk Assessment of Minor Enterprises

Huichao Mi
College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, Henan, China

Abstract

Under the influence of COVID-19, minor enterprises, especially the manufacturing industry, are facing greater financial pressure and the possibility of non-performing loans is increasing. It is very important for financial institutions to reduce financial risks while providing financial support for minor enterprises to promote industrial development and economic recovery. In order to understand the function of machine learning algorithms in predicting enterprise credit risk, the research designs five models, including Logistic Regression, Decision Tree, Naïve Bayesian, Support Vector Machine and Deep Neural Network, and adopts SMOTE and Undersampling to process imbalanced data. Experiments show that machine learning algorithms have high accuracy for both large-scale data and small-scale data.

Keywords*: credit risks, minor enterprises, Logistic Regression, Decision Tree algorithm, Naïve Bayesian, Support Vector Machine, Deep Neural Networks, SMOTE, Undersampling*

I. Introduction

Minor enterprises, especially the manufacturing industry, are an important force to promote economic development and solve people's employment problems. The development of minor enterprises is related to the development of the national economy and social stability, but the financing difficulties of minor enterprises restrict their rapid development. Compared with large enterprises, it is more difficult for Minor enterprises to obtain funds from the stock market, and their main source of funds is still financial institutions. However, due to the poor anti-risk ability of minor enterprises, the proportion of non-performing loans is higher. In 2020, affected by the epidemic, some Minor enterprises, especially those in manufacturing, wholesale and retail industries as well as those mainly in foreign trade orders, will face great financial pressure and may accumulate credit risks. According to data released by the China Banking and Insurance Regulatory Commission, the non-performing loan ratio of minor enterprises was 5.9% in 2018, while in the first quarter of 2020, the balance of non-performing loans of small and micro enterprises accounted for about 33.3% of the balance of non-performing loans of commercial banks [1]. In order to reduce financial risks brought by minor enterprises, financial institutions need to accurately predict the capital and operation of minor enterprises and evaluate their credit risks while providing financial support for them.

In recent years, with the rapid development of artificial intelligence and machine learning, its application and influence have become wider and wider, and the financial industry has become one of the earliest and most comprehensive industries in the application of artificial intelligence technology. Machine learning has long been used to assess personal and corporate credit. In the early days, people used statistical methods, such as logistic regression analysis, linear discriminant analysis, etc. Although these methods are simple to implement and have good interpretability, the dataset need to fit the distribution hypothesis, has poor applicability [2]. Wang et al. combined unsupervised learning with supervised learning to construct a combinational learning method for credit scoring of financial institutions [3]. But so far, there are few researches on applying machine learning algorithms to credit risk assessment of minor enterprises.

In order to understand the role of machine learning algorithms in SME credit risk assessment, this paper uses a

ISSN: 0010-8189

variety of machine learning algorithms to analyze business data, including Logic Regression (LR), Decision Tree Algorithm (DT), Naïve Bayesian (NB), Support Vector Machine (SVM) and Deep Neural Network (DNN). The experiment compares the evaluation results and judges the prediction accuracy and differences of different machine learning algorithms in risk assessment applications, so as to provide auxiliary decisions for financial institutions to avoid non-performing loan risks.

II. Research Status of Machine Learning Algorithms in the Financial Field

Artificial intelligence and big data are seen as the core of the Fourth Industrial Revolution. More and more credit decision-making processes are fully automated, improving overall decision-making efficiency. Jung used text mining and social network analysis technology to analyze the text data in the news from 2017 to 2019, and identified the current application mode of artificial intelligence in the financial field by mining the connection between keywords. Since financial products do not have obvious characteristics of profit and loss signs, He believes that financial institutions can apply clustering algorithms in scientific decision-making by matching various financial big data.

With the in-depth research on artificial intelligence technology, various artificial intelligence applications in the financial field are increasing. Naidu proposed a model based on Artificial Neural Network (ANN) and random forest, which innovated bankruptcy prediction in financial decision-making [6]. There are also researchers who use graph computing principles to reduce fraud detection in financial crimes [7]. Li divides the application of artificial intelligence and machine learning in the financial field into four areas: customer-oriented applications, management-level applications, market transactions and portfolio management, and regulatory compliance management [8].

The close integration of artificial intelligence and the financial field can improve the ability to prevent financial risks, guard the financial security system, and is changing people's production and lifestyles, providing more accurate and insightful data analysis for more effective risk management and investment portfolio to improve work efficiency and accuracy.

III. Machine Learning Algorithm

According to different learning methods, machine learning algorithms can be divided into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Among them, the difference between supervised learning algorithm, unsupervised learning algorithm and semi-supervised learning algorithm is whether the data used has been marked in advance.

3.1 Supervised Learning Algorithm

Training data set used by supervised learning algorithm must be all labeled in advance, and the machine learning algorithm uses the labeled data as experience to learn new knowledge. Common supervised learning algorithms include: k-Nearest Neighbor, Linear Regression, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest, Naive Bayes, etc. Most algorithms can solve both classification problems and regression problems, such as support vector machines, decision trees, naive Bayes, etc., but logistic regression can only be used to solve classification problems, while linear regression is specifically used to solve regression problems.

For enterprise credit risk assessment, classification problems include predictions such as risk level, whether the enterprise will go bankrupt, etc. If you need to predict the risk score or bankruptcy probability, you need to use regression algorithms to solve it.

LR is an extended usage of Linear Regression. In short, LR is to add a logistic model on the basis of Linear Regression, so that the model used to predict continuous feature becomes a predictive discrete feature. LR does not

directly model the target feature, but models the probability of the target attribute belonging to a certain category. For example, when conducting credit risk assessment, we can predict the probability of user default, which is recorded as $P(\text{default})$. If $P(\text{default}) \geq 0.5$, then it can be considered that the user will default. Otherwise, we can predict that the user will not default. Of course, if the institution wants to be more stringent in the assessment of default risk, the probability threshold can be lower. For example, if the default probability is less than 0.1, it is considered that there is risk of default. The formula of the LR model is as follows:

$$L(\text{default}) = \begin{cases} 1 & P(\text{default}) > 0.5 \\ 0 & P(\text{default}) \leq 0.5 \end{cases} \tag{1}$$

Different from other algorithms that require a large amount of training data to get a high-performance model, SVM are considered to be one of the most robust and accurate methods. Even with only a dozen data samples, SVM can get a highly accurate prediction model. At the same time, it is not sensitive to the dimension of the feature.

Although SVM can solve both classification problems and regression problems, it is mainly used in classification problems. Its essence is to place data in a high-dimensional space and find a separating hyperplane to separate two different categories as much as possible (positive Class and negative class), and make the distance of each class to the hyperplane as far as possible. The formula of separating hyperplane is as follows:

$$W^T \cdot X + b = 0 \tag{2}$$

where X is the feature vector, W is the normal vector of the SVM hyperplane, which determines the direction of the hyperplane, and b represents the offset, which determines the position of the hyperplane.

In order to improve the generalization performance of the model, SVM introduces slack variable, allowing some points to lie in the middle of the supporting hyperplane, and even be misclassified. For nonlinear separable data, SVM can use a kernel function to transform low-dimensional indivisible data into high-dimensional separable data. The formula of separating hyperplane is as follows:

$$\begin{gathered} \min_{W,b} \frac{1}{2}||W||^2 + C\sum_1^n \xi_i \\ \text{s.t. } y_i(W^T \cdot X_i + b) \geq 1 - \xi_i, \text{i} = 1,2, \ldots \ldots, \text{n} \end{gathered} \tag{3}$$

where C is the penalty factor, ξ_i is the slack variable C is used to balance the accuracy and generalization of the model.

Decision tree (DT) is a tree structure classification model or regression model, which is composed of nodes and edges. DT divides the data set by continuously constructing dividing conditions until the stopping conditions are met. There are two types of nodes in DT: internal nodes and leaf nodes. Taking the classification model as an example, the internal nodes give the rules for dividing the data set, and the leaf nodes are the final classification results. DT can be regarded as a set of rules composed of "IF-THEN". The path from the root node to the leaf node constitutes the classification rules of the model. For any data, there is one and only one rule corresponding to it.

Naive Bayes is a classification algorithm based on Bayes theorem and feature independence hypothesis. It is one of the most common classification algorithms. Compared with DT, NB is based on probability and statistics theory, and the classification performance is more stable. It calculates the posterior probability of each category to achieve classification. The NB algorithm is a simplification of Bayesian method, which assumes that the features are independent each other and have the same impact on the target feature. The core of NB algorithm is the assumption of feature independence. Using this assumption, the conditional probability formula is as follows:

$$P(X|y = C_k) = \prod_1^m P(x_i|y = C_k) \tag{4}$$

Where X is feature set, composed of m features, $X = \{x_1, x_2, \dots \dots, x_m\}$, is the kth category. Using this formula, the NB model can be learned and constructed, and y with the largest posterior probability under X features is selected as the prediction category. Although the NB algorithm is simple, in actual situations, features often have a certain correlation, and the assumption of feature independence is difficult to meet.

3.2 Unsupervised machine learning

Unsupervised learning uses unlabeled data as training data, observes the internal relationships of the data, discovers the rules, divides the data set into different class, and completes the labeling of the data set. Clustering algorithm is a typical unsupervised learning algorithm.

The clustering algorithm adopts the idea of "things are clustered together", which divides data with higher similarity into the same class, which is called "cluster". The goal of clustering is to divide the data set into different clusters, so that the data similarity within the cluster is as high as possible, and the data similarity between the clusters is as low as possible.

Clustering algorithm divides dataset D into k different clusters $\{C_1, C_2, \cdots, C_k\}$, the form of which is described as follows:

$$\bigcup_{1}^{k} C_i = D, \text{where } C_i \neq \emptyset \text{ and } C_i \in D$$

$$\forall C_i, C_j, C_i \cap C_j = \emptyset, \text{where } i, j = 1,2, \cdots k \text{ and } C_i, C_j \in D$$

The k-Means algorithm is the most famous clustering algorithm. By calculating the distance between the data point and the center of each cluster, the data points are classified to the nearest cluster, and then the center of the new cluster is recalculated. The algorithm continuously adjusts the cluster until the center of the cluster no longer changes.

3.3 Semi-Supervised Learning

Semi-Supervised Learning is a type of algorithm between supervised learning and unsupervised learning. Different from the fully labeled data set used in supervised learning and the unlabeled data set used in unsupervised learning, the dataset used in semi-supervised learning consists of a small amount of labeled data and a large amount of unlabeled data. The main idea of the semi-supervised machine learning algorithm is to automatically label unknown data with the help of a small amount of labeled data, and continuously expand the labeled training set to obtain a high-performance prediction model. Since semi-supervised learning only requires a small amount of labeled data, it can avoid the problem that the label cost of supervised learning is too high. Meanwhile, it uses labeled data to guide the label of unknown data, which reduces the disadvantage of low performance of unsupervised learning model and has high accuracy.

Semi-supervised learning was proposed in 1992, Merz combined semi-supervised learning with Adaptive Resonance Theory (ART) to solve classification problems [9]. Semi-supervised learning relies on model assumptions. If the hypothetical relationship between labeled data and unlabeled data is met, the unlabeled data can promote the construction of the final prediction model, otherwise it will reduce the accuracy of the model.

3.4 Artificial Neural Networks

Artificial Neural Networks (ANN) is an important method in machine learning. The simplest ANN model is single-layer perceptron that only contains an input layer and an output layer. When more processing layers are added

between the input layer and the output layer, single-layer perceptron becomes multi-layer perceptron, and the processing layers are also called the hidden layer. Multi-layer perceptron containing multiple hidden layers are also called Deep Neural Networks (DNN), which are extended on the basis of single-layer perceptron. DNN consists of three parts: input layer, hidden layer, and output layer. The hidden layer can have multiple layers, and each layer of hidden layer can contain multiple neurons. The neurons are fully connected, that is, each neuron of the layer is connected to each neuron of the next layer. When the output layer has multiple neurons, DNN can solve the multi-classification problem. Hinton believes that neural networks with multiple hidden layers have excellent learning capabilities [10].

The input layer of DNN sends the preprocessed data to the hidden layer after linear calculation. Each hidden layer uses the activation function to process the data. In this process, DNN continuously adjusts the model parameters until the loss function value of the model meets the performance requirements.

IV. Application of Machine Learning in Credit Risk Assessment

4.1 LR and credit risk assessment

Logistic regression is the most commonly used algorithm for classification problems, especially binary classification problems. In the financial field, logistic regression is mainly used in credit risk assessment, credit strategy research, venture capital analysis, bank credit decision-making, etc. Wang et al. used logistic regression to analyze the annual reports of 88 listed companies in 2004 and found that there was a correlation between the release of restatement announcements and management changes [11]. Schreiner et al. used a logistic regression model to build a scorecard and scored 39,956 samples of Bolivian microfinance creditors [12]. Fantazzini et al. compared the effects of the random survival forest model and the standard Logit model in the credit risk assessment of minor enterprises, and found that the simple Logit model had better performance in the out-of-sample prediction [13]. Hong et al. identified fraudulent behaviors in financial reports of listed companies by establishing backward step-by-step LR model [14].

4.2 DT and credit risk assessment

Decision tree is widely used in economy, industrial automation, finance and other fields. In 1985, Makowski used decision tree algorithm in personal credit evaluation to divide users into different groups according to different values of feature vectors, so as to achieve classification [15]. Li et al. used classification trees to classify credit risks of loans and compared them with Linear Discrimination Analysis (LDA) algorithm. Experiments showed that classification trees had high accuracy [16]. With the help of blockchain, decision tree and other technologies, Zhao designed the credit evaluation process and established the personal credit evaluation technology. Experiments show that this technology can effectively improve the transparency of personal credit information in Internet finance, and provide a new solution for the intelligent transformation and upgrading of Internet finance [17]. Zheng et al. used decision tree, logistic regression and kNN algorithm to build credit risk assessment model of listed companies. Through comparison, they found that the decision tree-based evaluation model had a higher overall judgment accuracy [18]. Ma et al. collected 26 dimensions of data from 1292 P2P companies in China, combined risk warning and machine learning, and constructed a risk warning model for P2P lending platforms. The model consists of a DT model, a NB model, and a SVM model [19]. The research results show that: registered capital, changes in business scope, platform background and other indicators have a greater impact on the accuracy of P2P risk prediction, and the DT model is the best model for risk early warning.

4.3 Clustering algorithm and credit risk assessment

In view of the new characteristics of financial big data, clustering algorithms have seen many new applications in credit risk assessment. Qian et al. used fuzzy clustering algorithm to assess the money laundering risks in 22 financial institutions, and classified these institutions into three categories: high, medium, and low. They found that

the anti-money laundering work in the banking industry is better than that in the securities industry and the insurance industry, and finally gave suggested measures based on the clustering results [20]. Zhao et al. used a multi-layer core set aggregation algorithm for financial data, clustered the volatility curve of stocks, discovered the inherent characteristics of stock volatility, and divided the stock sectors [21]. Fan proposed an improved k-Means algorithm to Segmentate bank customers [22]. The algorithm used the effective index method to dynamically adjust the initial cluster k, and used the adaptive optimal density radius method to determine the center of the cluster, which reducing the parameter pair The impact of algorithms. Tang construct data modeling and portrait of customer behaviors on financial e-commerce platforms, then clustered customer portraits using k-Means algorithm, hierarchized customers, and finally formulated personalized marketing plans for all levels customers [23]. Zeng proposed an improved BIRCH algorithm based on k-Mediods. Based on nearly 2 million transaction records of a securities company in the past 10 years, he clustered 662 customers of a securities company, and identified six types of customers with interactive attributes such as risk, attention, value, and maturity. The refinement of customers helps financial companies establish a reliable risk management mechanism [24].

4.4 Semi-supervised learning algorithm and credit risk assessment

With the development of Internet finance, the amount of data in the financial industry has increased exponentially, and the labor cost of labeling data is too high. However, the financial industry has high requirements for the accuracy of the model, and the accuracy of the model obtained by clustering using unlabeled data is difficult to meet the requirements. In addition, for issues in the financial field, the participation of senior financial analysts is often required to achieve relatively accurate data annotation and meet business requirements. This not only increases the cost of annotation, but also increases the time cost. Therefore, the use of semi-supervised learning algorithm can effectively reduce labor costs and obtain a highly reliable prediction model.

Considering the lack of data credit labels on online lending platforms and the difficulty in obtaining authoritative results, Ye proposed a semi-supervised learning algorithm based on BP neural network and collaborative training to establish a credit evaluation model for Internet financial online lending platforms. Experiments showed that semi-supervised learning method for risk assessment has higher performance than supervised learning model [25]. Aiming at the current risks in P2P lending, Zhou et al. proposed a speculative decision analysis (IDEA) based on a graph semi-supervised learning algorithm. The algorithm first constructs an investor-loan network graph, and then measures the bad debt risk of loans through investment behavior information. Finally, it uses the graph semi-supervised learning algorithm to evaluate the risk of loan flow mark, generate the best loan candidate set, and help investors make decisions [26]. Zhang established an efficient P2P online lending platform risk assessment model (LapSVM-PT). First, he analyzed public opinion on the platform, obtained data distribution in different dimensions, and then established feature vectors. Finally, LapSVM is used to evaluate the risk of the platform to determine whether it has the risk of bankruptcy [27]. Wen combined Deep Belief Network (DBN) with the isolated forest algorithm, and proposed a semi-supervised algorithm based on DBN-iForest, which is used to construct a credit default recognition model. The algorithm takes advantage of the feature learning ability of DBN and iForest's ability to recognize abnormal behaviors, uses simulated annealing algorithm and particle swarm optimization algorithm to achieve the optimization of the main parameters of the algorithm. It can identify fraudulent transactions from high-dimensional unbalanced credit financial transaction data [28].

4.5 ANN and semi-supervised learning

Due to the increasingly large and complex financial data, traditional machine learning algorithms are gradually unable to meet the needs, and neural networks have begun to be applied in the financial field. Bankruptcy prediction, debt risk assessment, and securities market applications are the three areas where neural networks are used more frequently. Compared with traditional statistical tools, ANN shows better generalization capabilities.

Tam used the BP network to predict the bankruptcy risk of the Texas bank in the United States. He increased the prior probability and misclassification cost during the training process, and compared it with DA, factor-logistic,

kNN and ID3. Experiments show that BP network is better than other methods in predicting accuracy. Meanwhile, researchers believe that neural network can be extended to other financial applications, especially those involving credit score and loan evaluation classification [29]. Alici constructed a three-layer neural network model to predict the bankruptcy of British companies. The model contains 28 inputs, 7 neurons in the hidden layer and 2 neurons in the output layer [30]. Beytollahi et al. predicted the price of the credit default swap (CDS) contracts of 125 companies from 2008 to 2015, and compared the predictive capabilities of various neural network models. The results show that the average predictive capability of Neural Network AutoregRessive with eXogenous input (NNARX) is higher than other Algorithms [31]. Wang et al. proposed a method based on convolutional neural network for customer credit risk assessment. They improved CNN to automatically extract features. The results show that the prediction effect of the convolutional neural network model for credit risk is better than logistic regression and random forest [32]. Li et al. used multi-layer perceptron and radial basis functions to evaluate the credit of minor enterprises in the P2P online lending platform and multi-layer perceptron can better predict defaulting enterprises [33].

V. Experimental Process and Result Analysis

5.1 Data preprocessing

This study uses public dataset downloaded from Kaggle. The Taiwan Economic Journal counts the business operations from 1999 to 2009, involving 346 manufacturing companies and 132 other types of companies. The dataset has 6819 records, including 96 characteristics such as ROA Operating Gross Margin, Realized Sales Gross Margin, operating profit rate, Continuous interest rate (after tax), Operating Expense Rate, Research and development expense rate, Cash flow rate, Bankruptcy. We need to establish a model to predict whether the enterprise is at risk of bankruptcy. There are only 220 records of bankrupt companies in the data set, accounting for 3.2%, and the distribution of the dataset is seriously unbalanced. In order to improve the accuracy of prediction, the dataset needs to be balanced. In this study, SMOTE and Undersampling were used to balance the dataset, and the efficiency of the algorithm was discussed in the case of different data volumes. The dataset obtained by the Undersampling method contains 500 pieces of data, and the dataset using the SMOTE method is 13,198 pieces.

The dataset includes 96 features. By analyzing the correlation with the target attributes and observing the distribution of the data, we finally selected 6 features to predict whether the company will go bankrupt, including ROA(C), Net Value Per Share (B), Debt ratio %, Total Asset Turnover, Cash/Total Assets, Equity to Liability. ROA and Total Asset Turnover reflect the operating efficiency of enterprise assets, which are important indicators to evaluate the asset management and profitability of an enterprise. Net Value Per Share and Debt ratio reflect the company's ability to repay debts. Cash/Total Assets is used to measure the liquidity of the company's assets, which can reflect the company's ability to pay current debts without relying on inventory sales and receivables. Equity to Liability calculates the proportion of corporate debt capital in total capital, which can reflect the degree to which debt capital is protected. This study will use these six characteristics to predict whether a company is at risk of bankruptcy.

5.2 Experimental process design

In order to study the prediction effects of machine learning algorithms in different scale datasets, this experiment uses two methods to process imbalanced data. After processing, two datasets of different scales are obtained, with 500 records and 13198 records respectively. 80% of the dataset is used as the training set and 20% is used as the test set. The same training set is used to construct LR, DT, NB, SVM and DNN prediction models to predict whether the company will go bankrupt; then use the same test set to test the effect of the model. Finally, the evaluation indicators of various models are calculated and compared, including F1-Score, Precision and Recall.

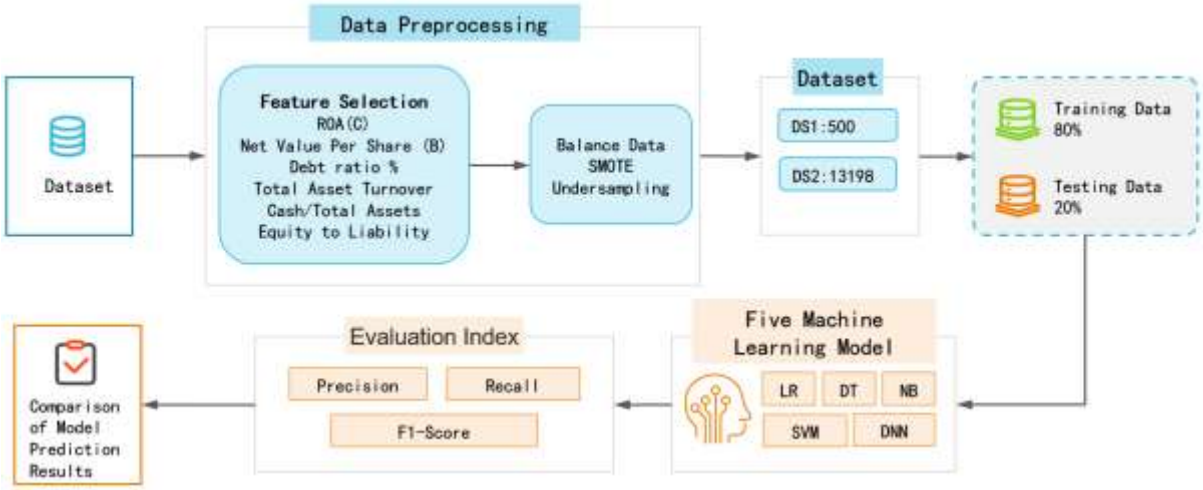


Fig 1: Comparison of the implementation process of machine learning prediction models

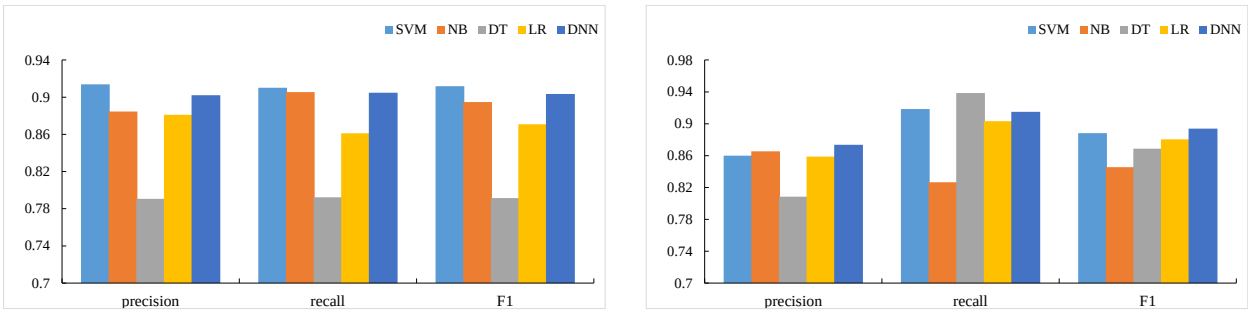
5.3 Discussion

This study uses Precision, Recall and F1-Score as evaluation indicators, uses two different sampling methods to process unbalanced data, and uses five different machine learning algorithms to build a predictive model. Finally, the model evaluation results are compared and analyzed.

The model evaluation results of the five models are shown in Table 1.

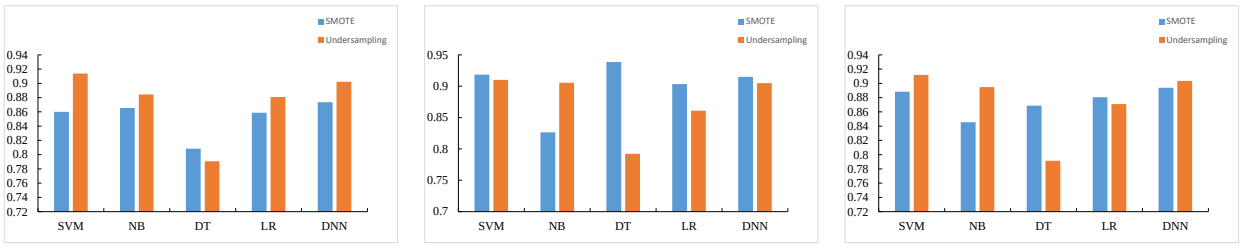
Table1 Model evaluation results of 5 models

Model Name	Model Evaluation Value	The method of sampling	
		SMOTE	Undersampling
LR	Precision	0.8588	0.881
	Recall	0.9033	0.8611
	F1-Score	0.8805	0.8709
DT	Precision	0.8084	0.7906
	Recall	0.9386	0.7923
	F1-Score	0.8687	0.7914
NB	Precision	0.8654	0.8845
	Recall	0.8265	0.9054
	F1-Score	0.8455	0.8948
SVM	Precision	0.86	0.9138
	Recall	0.9185	0.9101
	F1-Score	0.8883	0.9119
DNN	Precision	0.8736	0.9022
	Recall	0.915	0.9048
	F1-Score	0.8938	0.9034



(a) Undersampling *(b) SMOTE*

Fig 2: Comparison of evaluation results of five models



(a) Precision *(b) Recall* *(c) F1-Score*

Fig 3: Comparison of evaluation results of different sampling methods

Table 1 shows the performance indicators of the five models in different datasets, and Fig 2 shows the comparison of the indicators of these models in the dataset obtained by using the same sampling. When Undersampling is used to process imbalanced data, the data volume of the dataset is small, and all evaluation indicators of SVM model are higher than those of the other four models. Precision, Recall and F1-score of SVM model are 0.9138, 0.9101 and 0.9119, respectively, indicating that SVM has the best performance in small-scale data sets. On the contrary, the performance of DT model is the worst, about 10% lower than other models. When SMOTE is used to process unbalanced data, the size of dataset increases, and the performance of both DT model and LR model increases. Among them, DT model has the highest Recall value, reaching 0.9386, which is better than the other four models.

Fig 3 shows the comparison of different sampling methods for the same model. Except DT model, the Precision of all the other models in SMOTE sampled dataset is lower than that of the Undersampling sampled data set. The performance of DT model in SMOTE data set has been greatly improved, but the Precision of DT model is still lower than that of the other four models. Except for NB model, the other four models have better Recall on the SMOTE sampled dataset, and DT model has more obvious increase. For F1-Score, the changes in NB model and DT model are more obvious. The F1-Score of DT model on the SMOTE sampled dataset is much higher than that of the Undersampling sampled dataset, while NB model performs better on Undersampling dataset. The performance of DNN model on two datasets is not very different, and the performance is good. No matter what kind of dataset, the prediction performance of the five models is good, their accuracies are more than 80%, and the accuracy of the DNN model is close to 90%.

The results show that the distribution of business data of minor enterprises is not uniform, and it needs to be balanced first, while the traditional machine learning method is easily affected by the sampling method. Compared with logistic regression, decision tree, Naive Bayes and SVM, DNN, an artificial neural network with multiple hidden layers, has excellent feature learning ability, which is beneficial to predict classification. Considering various evaluation indexes of models, the research believes that machine learning algorithm is helpful for the credit risk assessment of minor enterprises, but traditional machine learning algorithm is greatly affected by dataset, while DNN performance is relatively stable.

ISSN: 0010-8189

VI. Conclusion

Aiming at the credit risk problems of minor enterprises, especially the manufacturing industry, this study used the operating data of about 500 companies in Taiwan region in 10 years, and selected the six most relevant features to predict the bankruptcy risk of enterprises. Experimental results show that the five machine learning algorithms can effectively predict the bankruptcy risk, and the DNN model has the most stable performance in different datasets. In the future research, we will continue to carry out the research on the credit risk level of minor enterprises. and study how to provide more detailed analysis for financial institutions to help them accurately identify high-potential enterprises and high-risk enterprises.

References

[1] Y. Y. Wu, “Risk aggregation and countermeasures of China's banking industry in the post-epidemic era”, The Chinese Banker, 69-71, 2020(09).
[2] Gang. W, Ma. J, “A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine”, Expert Systems with Application,39(5), p.5325-5331, 2012.
[3] Wang B, Ning L, Kong Y, “Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment”, Expert Systems with Applications, 128(AUG). 301-315, 2019.
[4] GO E J, MOON J, KIM J, “Analysis of the Current and Future of the Artificial Intelligence in Financial Industry with Big Data Techniques”, GLOBAL BUSINESS FINANCE REVIEW, 25(1): 102-117, 2020.
[5] Da 'an He, “Financial Big Data and Big Data Finance”, Academic Monthly, 51(12). 33-41, 2019.
[6] Naidu. G. P, Govinda K, “Bankruptcy prediction using neural networks”, 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018.
[7] Kurshan E, Shen H, Yu H, “Financial Crime & Fraud Detection Using Graph Computing: Application Considerations & Outlook”, 2020 Second International Conference on Transdisciplinary AI (TransAI), 2020.
[8] LI. C, “The Application of Artificial Intelligence and Machine Learning in Financial Stability: 2020 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy”, SPIoT-2020. Advances in Intelligent Systems and Computing (AISC 1282), 2021.
[9] Merz. C. J, Clair D, Bond W E, “SeMi-supervised adaptive resonance theory (SMART2)”, International Joint Conference on Neural Networks. IEEE, 1992.
[10] Hinton G E, Salakhutdinov, et al, “Reducing the Dimensionality of Data with Neural Networks”, Science, 2006.
[11] Zexia Wang, Mingming Li, Bing Xie, “Research on the effectiveness of financial statement restatement announcement and corporate governance”, Finance and Accounting Monthly, 2009(24):15-18.
[12] Schreiner, Mark, “A Scoring Model of the Risk of Costly Arrears at a Microfinance Lender in Bolivia”, Journal of Microfinance. 6, 2001.
[13] Dean. Fantazzini, Silvia. Figini, “Random Survival Forests Models for SME Credit Risk Measurement”, Methodology & Computing in Applied Probability, 2009.
[14] W. Z. Hong, X. X. Wang, H. Q. Feng, “Research on Identification of Fraud in Financial Reporting of Listed Companies Based on Logistic Regression Model”, Chinese Journal of Management Science, 22(S1). 351-356, 2014.
[15] Makowski. P. (1985), “Credit scoring branches out”, Credit World. 75. 30-37.
[16] Y. S. Li, W. Zhang, “How to apply classification tree to five loan classification in Commercial Banks of China”, Journal of Systems Engineering, 16(4). 282-288,2001.
[17] Y. Zhao, “Research on personal credit evaluation of internet finance based on blockchain and decision tree algorithm”, J Wireless Com Network. 213, 2020.
[18] Y. F. Zheng, J. Xu, “Empirical Research on the Credit Risk Model of Listed Companies Based on Decision Tree”, Communication of Finance and Accounting. 145-147, 2012(05).

[19] L. Y. Ma, Y. N. Wang, C. M. Ren, et al. “Early Warning for Internet Finance Industry Risk: An Empirical Investigation of the P2P Companies in the Coastal Regions of China", Journal of Coastal Research 106(sp1), 295-299, (10 July 2020)
[20] H. W. Qian, Peng. X, “the Application of Fuzzy Clustering Analysis to the Money Laundering Risk Assessment of Financial Institutions”, West China Finance. 93-96, 2014(08).
[21] T. C. Zhao, Ma. R. N, Q. Zhang, “A multilevel aggregation algorithm for financial data based on functional data analysis”, Mathematical Modeling and Its Applications, 9(03). 40-46,2020.
[22] Fan. N, “Simulation Study on Commercial Bank Customer Segmentation on K-means Clustering Algorithm”, Computer Simulation, 28(03). 369-372, 2011.
[23] Tang. C. H, “Research of Stock Clustering for Financial Knowledge Service”, Harbin Institute of Technology, 2017.
[24] X. D. Zeng, “A Big Data Clustering Method Based on K-Mediods Improved BIRCH”, Yunnan University of Finance and Economics, 2015.
[25] Ye. X. J, “A Trust Evaluation Model towards Internet-based Lending Platform Based on BP Neural Network and Semi-supervised Learning”, Zhejiang University, 2015.
[26] Y. H. Zhou, Y. Z. Zhang, Mi. J. H, “IDEA: An Investment Decision Analysis Algorithm for P2P Lending”, Computer Systems & Applications, 25(09). 200-206, 2016.
[27] X. M. Zhang, “P2P platform risk assessment model based on public opinion analysis and text theme”, Zhejiang University, 2018.
[28] L. N. Wen, “Research on credit default identification method based on deep learning”, Beijing University of Technology, 2019.
[29] Tam K Y, “Neural Network Models and the Prediction of Bank Bankruptcy”, Omega, 19(5). 429-445, 1991.
[30] Alici, “Neural networks in corporate failure prediction”, University of Exeter, 1996.
[31] Beytollahi A, Zeinali H, “Comparing Prediction Power of Artificial Neural Networks Compound Models in Predicting Credit Default Swap Prices through Black–Scholes–Merton Model”, Iranian Journal of Management Studies. 13, 2020.
[32] D. Zhang, S. Lou, “The application research of neural network and BP algorithm in stock price pattern classification and prediction – ScienceDirect”, Future Generation Computer Systems, 115. 872-879, 2021.
[33] S. J. Li, Pan. Y. H, “Application of neural network technology in credit evaluation of P2P network credit SMEs”, Productivity Research. 14-22+161, 2019(05).