A Comparative Study of Machine Learning Algorithms in the Prediction of Personal Loan Defaults

Silu Zhang
Faculty of Business, Macau University of Science and Technology, 999078, Macau, China
1148251688@qq.com

ABSTRACT

This article focuses on the comparative study of the application of machine learning algorithms in the field of personal loan default prediction. This paper articulates the critical role of predicting personal loan defaults and systematically dissects the mathematical principles and derivation logic of mainstream machine learning algorithms. It conducts a rigorous comparative analysis focusing on their theoretical foundations, optimization objectives, and distinct characteristics in handling financial data. This paper explores the considerations of algorithms in the practical application of personal loan default prediction, including data quality, model interpretability, and business demand matching degree, etc. Analyze the challenges currently faced by research and practice, such as data privacy and security, algorithmic bias, etc., and propose corresponding countermeasures. It aims to provide theoretical references for financial institutions to rationally select and apply machine learning algorithms in the prediction of personal loan defaults, and promote the improvement of financial risk management levels.

Keywords: Machine learning algorithm; Prediction of personal loan default; Comparative study; Financial risk management.

1. INTRODUCTION

In the financial sector, the scale of personal loan business has shown a continuous and significant expansion trend. With the development of the economy and the continuous opening up of the financial market, the demand for personal loans is growing increasingly strong, covering multiple fields such as housing loans, auto loans, and consumer loans. The vigorous development of this business not only brings considerable profits to financial institutions and becomes an important profit growth point for them, but also comes with the default risk that cannot be ignored. Default risk is like a "time bomb" hidden in financial business. Once it erupts, it may cause huge economic losses to financial institutions, seriously affect their asset quality and profitability, and even trigger systemic financial risks [1]. Therefore, accurately predicting the default situation of personal loans is a key link for financial institutions to control risks and ensure asset quality, which is of vital significance.

Traditional prediction methods have dominated the prediction of personal loan defaults for a relatively long time. Most of these methods are based on statistical models, such as logistic regression and linear discriminant analysis. These models demonstrate certain effectiveness when dealing with simple linear relationships and smaller-scale data, and can provide basic default risk assessment for financial institutions. However, with the increasing complexity of the financial market and the explosive growth of data volume, the limitations of traditional prediction methods have gradually become prominent. In dealing with complex nonlinear relationships, traditional statistical models often fall short. The default risk of personal loans in reality is intertwined with numerous factors, presenting complex nonlinear characteristics. For instance, the income level of borrowers and the probability of default are not in a simple linear relationship. Low-income groups may default due to insufficient repayment ability, but high-income groups may also default due to excessive consumption or investment failure and other reasons. The default situation of middle-income groups is even more complex. Traditional models have difficulty accurately capturing such complex nonlinear relationships, which leads to deviations in the prediction results [2].

Traditional methods also face many challenges in handling large-scale data. With the development of fintech, the dimensions of data that financial institutions can collect are constantly increasing. Besides traditional financial data and credit history data, it also includes social network data, consumer behavior data, etc. These data are large in scale and complex in structure. Traditional statistical models have low computational efficiency when processing these data and find it difficult to mine the deep-seated information hidden in the data, thus failing to fully leverage the value of the data

[3].

With the rapid development of artificial intelligence technology, machine learning algorithms, with their powerful data processing capabilities and pattern recognition capabilities, have gradually emerged in the field of personal loan default prediction. Machine learning algorithms can automatically learn patterns and rules from large-scale and complex data without the need for manual setting of complex rules and assumptions, and can better adapt to the changes and complexity of data. There are significant differences among various machine learning algorithms in terms of principle, performance and applicable scenarios. For instance, the decision tree algorithm makes decisions by constructing a tree-like structure, which is intuitive, easy to understand and interpret, and is suitable for handling discrete data. The support vector machine algorithm achieves classification by finding the optimal hyperplane and has advantages when dealing with high-dimensional data and small sample data. Neural network algorithms can simulate the neuronal structure of the human brain and have a powerful nonlinear fitting ability, making them suitable for handling complex nonlinear relationships [4].

By comparing and studying the application of these machine learning algorithms in the prediction of personal loan defaults, and deeply analyzing their performance differences and applicable scenarios, it is helpful for financial institutions to select appropriate algorithms for default prediction based on their own business needs, data characteristics and risk preferences, thereby improving the accuracy of prediction and risk management level, and achieving a balance between risk and return [5].

2. THE SIGNIFICANCE OF PREDICTING PERSONAL LOAN DEFAULTS

2.1 Risk Control of Financial Institutions

Personal loan default is like a sudden storm for financial institutions, which will bring direct and serious economic losses. When borrowers fail to repay the principal and interest of their loans on time, financial institutions will not only lose the expected interest income but also face the risk of not being able to recover the principal of the loans. Such losses will directly affect the asset quality of financial institutions, leading to an increase in the non-performing loan ratio and, in turn, their profitability and market reputation. For instance, during the global financial crisis in 2008, the default of a large number of personal housing loans led to a sharp deterioration in the asset quality of many financial institutions, and some even faced the crisis of bankruptcy and closure [6].

By accurately predicting the default risk of personal loans, financial institutions can take a series of effective measures in advance to reduce default losses. On the one hand, financial institutions can adjust the loan amount based on the prediction results. For borrowers with a higher risk of default, the loan amount can be appropriately reduced to minimize potential loss exposure. On the other hand, financial institutions can adjust loan interest rates and charge higher rates to high-risk borrowers to compensate for the possible default risks they may face. In addition, financial institutions can also require high-risk borrowers to provide more guarantees, such as collateral or guarantors, to enhance the security of loans [7].

Accurate prediction of personal loan default risks also helps financial institutions rationally allocate credit resources. Credit resources are one of the core resources of financial institutions. How to efficiently allocate them to customers with lower default risks is the key for financial institutions to improve the efficiency of fund utilization and competitiveness. Through default prediction, financial institutions can screen out high-quality customers with good credit standing and strong repayment ability, and allocate more credit resources to these customers, thereby improving the efficiency of fund utilization and increasing profits. At the same time, this also helps financial institutions optimize their customer structure, reduce overall default risks, and enhance their competitiveness in the market [8].

2.2 Stability of the Financial Market

Personal loan defaults not only affect individual financial institutions but may also have an impact on the stability of the entire financial market. A large number of personal loan defaults may trigger a chain reaction, leading to a decline in the asset quality of financial institutions and subsequently affecting the stability of the financial system [9]. Accurate prediction of personal loan defaults helps regulatory authorities promptly grasp the market risk situation, take corresponding regulatory measures, and maintain the stable operation of the financial market.

2.3 Improve the Personal Credit System

The prediction of personal loan defaults relies on personal credit information, and the prediction results, in turn, promote

the improvement of the personal credit system [10]. Accurate prediction of default situations can motivate individuals to attach importance to their credit records, repay on time and form good credit habits. At the same time, it provides more accurate data support for credit rating agencies, promotes the development of the personal credit rating system, and enhances the reference value of credit information in financial transactions.

3. COMMON TYPES AND CHARACTERISTICS OF MACHINE LEARNING ALGORITHMS

3.1 Supervised Learning Algorithm

Supervised learning involves training with labeled data to enable the model to make predictions about new data [11]. In personal loan default prediction, labeled data refers to data that includes the customer's historical loan information and the outcome of whether there is a default. Common supervised learning algorithms include logistic regression, decision trees, support vector machines, etc. Logistic regression, by establishing a linear regression model, maps the prediction results to probability values to determine the possibility of customer default. It is characterized by simplicity, ease of understanding, and high computational efficiency, but its ability to handle complex nonlinear relationships is limited. Decision trees classify data through a series of rules, are easy to understand and interpret, can handle multiple types of data, but may overfit and be sensitive to data noise. Support vector machines achieve classification by seeking the optimal separation hyperplane and perform well in dealing with high-dimensional data and small sample data, but the model selection and parameter adjustment are relatively complex.

3.2 Unsupervised Learning Algorithm

Unsupervised learning processes unlabeled data, aiming to discover potential structures and patterns within the data. In the prediction of personal loan defaults, it can be used for customer segmentation, dividing customers with similar characteristics into a group and formulating differentiated credit strategies for different groups of customers. Common unsupervised learning algorithms include clustering algorithms, such as K-means clustering, which divides data into K clusters by calculating the distances between data points. It is simple and fast, but the number of clusters needs to be specified in advance and it is sensitive to the initial values. Hierarchical clustering forms a hierarchical structure by continuously merging or splitting clusters without the need to specify the number of clusters in advance, but it has a relatively high computational complexity [12].

3.3 Ensemble Learning Algorithm

Ensemble learning improves model performance by combining multiple base learners. Common ensemble learning algorithms include random forests and gradient boosting trees. A random forest is composed of multiple decision trees. By randomly selecting samples and features to construct the decision trees and ultimately voting to determine the prediction results, it can effectively reduce the risk of overfitting and improve the generalization ability of the model, but the interpretability of the model is relatively poor. Gradient boost trees gradually optimize the model through an iterative approach. In each iteration, a new decision tree is established by reducing the residual direction of the previous iteration. It can handle complex nonlinear relationships and has high prediction accuracy, but it takes a long time to train and is sensitive to parameters.

4. COMPARATIVE ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS IN PREDICTING PERSONAL LOAN DEFAULTS

4.1 Compare from the Perspective of Theoretical Basis

Different machine learning algorithms are based on different mathematical theories and assumptions. Logistic regression is based on linear regression and maximum likelihood estimation. It assumes the data follows a Bernoulli distribution. The algorithm maps the linear combination of features into a probability interval $(0,1)$ using the Sigmoid function, which serves as the methodological basis for transforming discrete labels into continuous probabilities:

$$P(Y=1 \mid X)=\frac{1}{1+e^{-(w^{T}x+b)}} \tag{1}$$

This derivation allows the model to predict the probability of default by optimizing the parameters w and b via Maximum Likelihood Estimation.

Decision trees, grounded in information theory, recursively partition the feature space to maximize node purity. The algorithm selects the optimal splitting feature by minimizing the Gini Impurity index, ensuring that each derivation step effectively reduces uncertainty:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{2}$$

Where p_k represents the probability of a sample belonging to class k. This mathematical criterion drives the construction of decision rules.

Support vector machines adhere to the principle of structural risk minimization. The methodological core involves solving a convex optimization problem to find the hyperplane that maximizes the geometric margin between default and non-default classes:

$$\min_{w,b} \frac{1}{2} \| w \|^2 \quad \text{s.t.} \quad y_i (w^T x_i + b) \geq 1 \tag{3}$$

This objective function mathematically guarantees the maximum separation margin, enhancing the model's generalization ability on small sample data.

Unsupervised learning algorithms such as clustering algorithms cluster similar data together based on similarity measures between data. Ensemble learning algorithms improve overall performance by combining multiple base learners and leveraging the diversity among them. The theoretical basis involves probability theory, statistics, and optimization theory, etc.

4.2 Compare from the Perspective of Characteristics

Supervised learning algorithms have clear prediction objectives, can be directly trained using labeled data, and the prediction results are interpretable. However, they have high requirements for data quality and labeling accuracy. Unsupervised learning algorithms do not require labeled data and can discover hidden patterns and structures in the data. However, their prediction targets are not clear and the results need to be interpreted in combination with business knowledge. Ensemble learning algorithms can integrate the advantages of various algorithms by combining multiple base learners, improving the stability and accuracy of the model. However, the model complexity increases, the consumption of computing resources is relatively large, and the interpretability is relatively weak. Different algorithms also vary in terms of data types, data scales, and feature dimensions. For instance, decision trees can handle multiple types of features, and support vector machines perform well in high-dimensional data.

4.3 Compare from the Perspective of Applicable Scenarios

Logistic regression is applicable to scenarios where the linear relationship of data features is strong and the interpretability of the model is highly demanded, such as small financial institutions or situations where risk control requirements are relatively simple. Decision trees are suitable for scenarios where data features are complex and intuitive decision rules are required, and they can help business personnel understand the factors influencing customer defaults. Support vector machines are suitable for scenarios with a small sample size, high feature dimensions, and high requirements for model accuracy, such as predicting loan defaults of some high-end customers. Unsupervised learning algorithms are applicable to scenarios such as customer segmentation and market analysis, providing a basis for financial institutions to formulate differentiated marketing and credit strategies. Ensemble learning algorithms are applicable to scenarios where high prediction accuracy is required, the data scale is large, and sufficient computing resources can be provided, such as comprehensive risk assessment of large financial institutions.

5. CONSIDERATIONS OF MACHINE LEARNING ALGORITHMS IN THE PRACTICAL APPLICATION OF PERSONAL LOAN DEFAULT PREDICTION

5.1 Data Quality

Data quality is a key factor influencing the prediction performance of machine learning algorithms. Data integrity requires complete data records without missing values; otherwise, it may lead to inaccurate model training. Accuracy requires that the data be true and reliable, with no error records. For instance, data such as customer income and credit

records must be accurate and error-free. Consistency requires that data remain consistent across different sources and time periods to avoid data conflicts. Timeliness requires that data be updated in a timely manner to reflect the latest credit status and repayment ability of customers. Financial institutions need to establish a complete data quality management system, clean, preprocess and monitor the data to ensure that the data quality meets the requirements of model training.

5.2 Model Interpretability

In the financial field, the interpretability of models is of vital importance. Financial institutions need to explain the model prediction results and decision-making basis to regulatory authorities, clients and internal management. Algorithms such as logistic regression and decision trees have strong interpretability and can clearly display the influencing factors of customer default and decision-making rules. However, some complex ensemble learning algorithms and deep learning algorithms have relatively weak interpretability. Therefore, techniques such as feature importance analysis and local interpretability methods need to be adopted to enhance the interpretability of the models. When choosing algorithms, financial institutions need to comprehensively consider prediction accuracy and interpretability. Under the premise of meeting business requirements, they should try to select algorithms with stronger interpretability.

5.3 Business Requirement Matching Degree

Different financial institutions have varying business scales, customer groups and risk preferences, and thus have different demands for predicting personal loan defaults. Small financial institutions may pay more attention to the simplicity, ease of use and cost-effectiveness of models, and choose simple algorithms such as logistic regression. Large financial institutions have complex business operations and a large number of clients, which have high requirements for prediction accuracy and comprehensiveness. They can choose ensemble learning algorithms or combine multiple algorithms to build complex models. At the same time, financial institutions need to take into account the costs of model updates and maintenance to ensure that the models can adapt to business development and market changes. When choosing an algorithm, it is necessary to fully assess the matching degree between the algorithm and business requirements, and select the algorithm that best suits one's own business development.

6. CHALLENGES AND COUNTERMEASURES CURRENTLY FACED BY RESEARCH AND PRACTICE

6.1 Data Privacy and Security

Personal loan data contains sensitive customer information such as ID numbers, income, and credit records. Data privacy and security are important issues. During the process of data collection, storage, transmission and use, there may be risks of data leakage, tampering and abuse. Financial institutions need to enhance data security management, adopt encryption technology to encrypt data, and establish a strict data access permission control mechanism to ensure that only authorized personnel can access and use the data. At the same time, comply with relevant laws and regulations, such as the Personal Information Protection Law, to protect customer data privacy.

6.2 Algorithmic Bias

Machine learning algorithms rely on historical data for training. If historical data contains biases, such as those related to gender, race, and region, the algorithms may inherit and amplify these biases, leading to unfair prediction results for certain customer groups. Financial institutions need to review and preprocess data to eliminate biased factors in it. During the model training and evaluation process, fairness indicators are used to assess the fairness of the algorithm, such as the prediction accuracy of different groups and the false alarm rate, etc. At the same time, by integrating business knowledge and expert experience, the algorithm prediction results are manually reviewed and adjusted to ensure the fairness and impartiality of the prediction results.

6.3 Model Update and Maintenance

The market environment and customer credit status are constantly changing, and machine learning models need to be updated and maintained regularly to ensure the accuracy of predictions. Financial institutions need to establish a model monitoring mechanism to monitor model performance indicators in real time, such as accuracy rate, recall rate, F1 value, etc. When the model performance declines or there are significant changes in the market environment, update and optimize the model in a timely manner. Meanwhile, record the process and results of model updates, and establish a model version management system to facilitate traceability and management.

7. CONCLUSION

Machine learning algorithms have significant application value in the prediction of personal loan defaults. Different algorithms vary in theoretical basis, characteristics and applicable scenarios. When choosing algorithms, financial institutions need to comprehensively consider factors such as data quality, model interpretability, and the degree of matching with business requirements. Current research and practice are confronted with challenges such as data privacy and security, algorithmic bias, and model update and maintenance, and corresponding countermeasures need to be adopted to address them. Looking forward, the evolution of Fintech will likely prioritize the integration of high-performance ensemble models with Explainable AI (XAI) techniques. Coupled with improvements in data quality, these advancements will drive the application of machine learning algorithms to be more robust and transparent, playing a greater role in the risk management of financial institutions and the stability of the financial market. Financial institutions should continuously pay attention to technological development trends, constantly optimize and improve prediction models, and enhance the accuracy of personal loan default prediction and risk management level.

REFERENCES

[1] Fan, S. (2023, January). Design and implementation of a personal loan default prediction platform based on Light GBM model. In 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA) (pp. 1232-1236). IEEE.

[2] Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P lending default prediction based on AI and statistical models. Entropy, 24(6), 801.

[3] Wang, Q. (2022). Research on the method of predicting consumer financial loan default based on the big data model. Wireless Communications and Mobile Computing, 2022(1), 3786707.

[4] Chouksey, A., Shovon, M. S. S., Tannier, N. R., Bhowmik, P. K., Hossain, M., Rahman, M. S., ... & Hossain, M. S. (2023). Machine Learning-Based Risk Prediction Model for Loan Applications: Enhancing Decision-Making and Default Prevention. Journal of Business and Management Studies, 5(6), 160-176.

[5] Guo, W., & Zhou, Z. Z. (2022). A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction. Journal of Forecasting, 41(6), 1248-1313.

[6] Hamdillah, D. N., Purwanto, B., & Ermawati, W. J. (2021). The effect of asset quality and financial performance on non-performing loans (NPLs) of rural banks in Indonesia.

[7] Scott, A. O., Amajuoyi, P., & Adeusi, K. B. (2024). Effective credit risk mitigation strategies: Solutions for reducing exposure in financial institutions. Magna Scientia Advanced Research and Reviews, 11(1), 198-211.

[8] Alvi, J., Arif, I., & Nizam, K. (2024). Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management. Heliyon, 10(21).

[9] R. Wu, "Dynamic Credit Risk Assessment Based on Multi-Task Deep Learning and Bayesian Optimization," 2024 3rd International Conference on Smart City Challenges & Outcomes for Urban Transformation (SCOUT), Bhubaneswar, India, 2024, pp. 162-168, doi: 10.1109/SCOUT64349.2024.00040.

[10] Guan, Y., Wang, Y., & Yang, Y. (2024). Personal Loan Default Prediction Analysis based on TabNet and Logistic Regression. Applied and Computational Engineering.

[11] Özkurt, C. (2024). Enhancing Financial Decision-Making: Predictive Modeling for Personal Loan Eligibility with Gradient Boosting, XGBoost, and AdaBoost. ADBA Information Technology and Publishing Limited Company.

[12] R. Wu, "Multivariate Financial Time Series Forecasting Model Based on Transformer Architecture," 2024 3rd International Conference on Smart City Challenges & Outcomes for Urban Transformation (SCOUT), Bhubaneswar, India, 2024, pp. 155-161, doi: 10.1109/SCOUT64349.2024.00039.