Financial Customer Behavior Prediction Based on Machine Learning: A Comprehensive Investigation

Xinyue Zhang*

Digital Economics, Minzu University of China, 100000 Beijing, China

Abstract. Predicting customer behavior has become a critical component in shaping effective financial strategies. As customers' expectations evolve and their behavior becomes increasingly complex, traditional methods struggle to keep up with the demands for accuracy and efficiency in analysis. This paper reviews the financial customer behavior prediction technology based on machine learning (ML), emphasizing its importance in the formulation of financial industry strategies. The paper first introduces how the machine learning is applied in financial customer behavior prediction, including data collection, preprocessing, feature extraction and model selection. Then, by comparing deep learning and traditional machine learning models, their applications and effects in customer churn and loan prediction are explored. The paper also discusses challenges such as model interpretability, data distribution differences and privacy protection, and looks forward to future research directions, such as integrating machine learning techniques, tools to improve model interpretability, and transfer learning strategies. Finally, the paper summarizes the positive impact of machine learning in financial customer behavior prediction.

1 Introduction

In the contemporary financial landscape, the financial sector is at the epicenter of economic activity, with customer behavior playing an essential role in shaping the strategies of financial institutions. Thus, the ability to accurately predict customer behavior has become paramount for institutions seeking to enhance customer engagement, optimize marketing strategies, and mitigate risks. The advent of big data, coupled with the proliferation of digital financial services, has ushered in a new era of customer behavior analysis. Meanwhile, the limitations of traditional financial and mathematical models in capturing these nuances are becoming increasingly apparent. Classic regression analyses and heuristic-based approaches often fall short in predicting the intricacies of customer behavior, particularly in the face of rapidly evolving market conditions and the growing complexity of financial products. The need for more sophisticated predictive methods has never been

* Corresponding author: 22010781@muc.edu.cn

more critical, prompting a shift towards Artificial Intelligence (AI) and Machine Learning (ML) techniques.

The significance of the finance customer behavior prediction lies in its potential to help financial institutions gain a competitive edge. Through prediction, financial institutions can better understand the needs of each customer, anticipate market trends, and take preemptive measures to prevent customer attrition. In addition, the predictive power of machine learning can also help financial institutions do a better job in credit scoring, fraud detection, and risk assessment, thereby making the financial system more stable and reducing the risks brought by uncertainties.

The journey of ML in financial customer behavior prediction has been marked by significant milestones. Early studies focused on leveraging logistic regression and decision trees to model customer behavior based on historical transaction data [1]. As the field evolved, more sophisticated algorithms such as random forests and support vector machines were introduced, offering improved accuracy and efficiency. The recent surge in research has been driven by the development of ensemble methods like XGBoost and LightGBM, which have demonstrated remarkable prowess in handling complex, high-dimensional data sets [2, 3]. The latest study, as exemplified by Romanenko et al. and Zuama et al., underscores the increasing sophistication of ML models in predicting financial customer behavior [4]. These studies highlight the importance of feature engineering, the role of advanced algorithms, and the impact of model evaluation metrics in enhancing predictive accuracy. Soon after, deep learning algorithms, especially neural network models, began to play a major role in predicting customer behavior for long-term deposits because they could process large-scale data and extract high-level feature representations. Now researchers have begun to explore model fusion technologies, such as Stacking, which combine the advantages of multiple machine learning models to optimize prediction results through meta-learners [5].

This study aims to review the application of machine learning in financial customer behavior prediction and explore how it can help financial institutions improve services, predict market trends, and reduce customer churn risks. The other parts of the article are structured as follows. The author will introduce and classify machine learning workflow and basic models for financial customer behavior prediction in Section 2. Progressing further, Section 3 will compare deep learning and traditional machine learning models and provide prospects for future development. Conclusively, Section 4 encapsulates the main findings of the article and summarizes the research impact and points out the deficiencies.

2 Method

2.1 The introduction of machine learning workflow

The development of machine learning models usually follows the workflow below: Firstly, collect relevant financial customer data, which may include transaction records, customer personal information, account balances, etc. The data source can be the bank's internal database, public data set, or real-time data obtained through API. Then the preprocessing is carried out on the collected data. For example, handle missing values, remove duplicate records, standardize data formats, and convert data types. At this stage, the data will be organized into a format suitable for model training. After the data is clear, it will extract features that help model learning from the original data and select a suitable machine learning algorithm to build a predictive model. In financial customer behavior prediction, commonly used algorithms include logistic regression, random forest, support vector machine, neural network, etc. After the that, it is necessary to train and test the model using

historical data and evaluate its performance to determine its predictive ability.

2.2 Churn prediction

2.2.1 Neural network

Neural networks, especially deep learning models, show great potential in customer churn prediction. Neural networks are able to learn and simulate complex nonlinear relationships by imitating the working mode of human brain neurons, which makes them more effective than traditional methods when large-scale data are processed. In customer churn prediction, Long Short-term Memory (LSTM) networks and Convolutional Neural Networks (CNN) are usually used. LSTM networks are particularly suitable for processing time series data, while CNN can effectively process data with spatial hierarchical structures. An input layer, a hidden layer, and an output layer consist of a typical neural network model. The hidden layer can contain multiple neurons, which are connected by weights. Data preprocessing usually includes normalization, missing value processing, feature encoding, etc. This method combines a hybrid model of LSTM and CNN to capture both the long-term dependencies of time series data and the local features of spatial data [6]. Meanwhile, it's very useful to use the attention mechanism to increase the model's attention to key characteristics, thereby improving prediction accuracy. Omer Faruk SEYMEN et al. use ordinary artificial neural networks (ANNs) and convolutional neural networks (CNNs) for churn prediction in the retail industry. They used an ANN model with three hidden layers of 200 neurons each, while designing a CNN model to predict customer churn by converting customer transactions and demographics into images. The CNN model shows higher accuracy than the traditional model in this task and does not require manual feature extraction [6].

2.2.2 Random forest

When predicting customer churn, random forest is also a commonly used method. Building decision trees and then integrating the results could greatly enhance the accuracy and robustness of predictions. A subset of features would be randomly selected by each tree to split during training, by which the diversity of the model could be increased, and the risk of overfitting also could be reduced. Random forest is made up of multiple decision trees, and each of the tree has a leaf node representing a classification result, and the final prediction result is generated by majority voting. Similar to neural networks, random forests also require data preprocessing, including processing missing values, feature selection, and data encoding [7]. The innovation of the random forest method is that it can help identify which features are most critical to customer churn prediction through feature importance scoring. It can also handle large amounts of data and has good prediction performance for unbalanced data sets. Theresa Gattermann-Itschert and Ulrich W. Thoonemann examine how machine learning models can be used to predict customer churn and take proactive customer retention measures in the non-contractual B2B (business-to-business) wholesale industry. The study trained a random forest model and tuned hyperparameters with grid search and 5-fold cross-validation, followed by multi-slicing and out-of-period testing to evaluate model performance. The results of field experiments show that the churn rate of target customers predicted based on the model is significantly reduced compared with the random selection of target customers [8].

2.3 Loan prediction

2.3.1 Random forest

Studies show that in loan prediction, random forests are also widely used due to their excellent performance on prediction and ability in generalization of data sets. Random forests improve prediction accuracy by integrating decision trees. And each decision tree is trained on a different subset of the data set, and the final prediction result is determined by majority voting or average. In loan prediction, the ratio of defaulting and non-defaulting customers may be unbalanced. Random forests can be combined with oversampling or undersampling techniques, such as the SMOTE algorithm, to deal with this imbalance and improve the prediction accuracy of defaulting customers. Meanwhile, random forests are combined with deep learning models, and random forests are used for feature selection or dimensionality reduction, and then the screened features are input into the deep learning model to improve prediction performance [9]. Not only that, but a dynamic weight mechanism is also introduced in random forests to assign different weights according to the prediction performance of each tree, making the final prediction result more dependent on trees with better performance. Mehul Madaan et al. used the publicly available Lending Club dataset in their study, which covers approximately 220,000 loans issued between 2007 and 2015. They trained two models, a decision tree and a random forest, to analyze the dataset and learn patterns from it, and then predict whether a new applicant will default. Finally, use performance metrics to evaluate the model. It could be concluded that in loan default prediction, the random forest model performs better and is more accurate, and the model can help banks identify borrowers with potential default risks, thereby reducing credit risk [10].

2.3.2 Gradient boosting machines

Gradient boosting machine is a powerful ensemble learning technique that improves the performance of the model and achieves more accurate predictions by continuously adding decision trees and training each tree on the residuals of all previous trees. In loan prediction, GBM can customize the loss function for different business needs. For example, a loss function can be designed to apply different penalty weights to different types of prediction errors (such as misclassifying good customers as high-risk customers) to make the model perform optimally in a specific business scenario [11]. At the same time, GBM can be extended to multi-task learning scenarios to simultaneously predict multiple related targets in loan prediction, such as loan default probability, default time, loan amount, etc. This multi-task learning method can help the model learn multiple related tasks under a unified framework. Instead of simply using a single feature to split the data, GBM considers the interaction between features. This approach can help the model capture more complex data patterns, thereby improving the accuracy of predictions. Mayank Anand et al. used loan datasets from multiple Internet sources such as Kaggle and trained multiple models including the Gradient Boosting Machine (GBM) to analyze the dataset and learn patterns from it, identifying potential problem customers in a large number of loan applications, providing a more effective basis for loan credit approval [12].

3 Discussion

3.1 Comparison

In the field of predicting financial consumer behavior, deep learning models can automatically learn complex feature representations from large amounts of data, without the need for manual feature engineering, compared to traditional machine learning models. This is especially important when dealing with unstructured data, because feature extraction of these data often requires a relatively high degree of abstraction. Because deep learning models are trained on large data sets, subtle patterns and trends in the data could be easily captured, thereby providing higher accuracy in predicting consumer behavior. But it also has drawbacks. The deep learning models' performance is highly dependent on the quality and quantity of training data. The model's prediction results will also be affected when the data set is incomplete. Not only that, with a large number of parameters, deep learning models may overfit the training data, resulting in a decrease in generalization ability on new data.

3.2 Limitations and challenges

From a macro perspective, limitations and challenges still existed when applying machine learning to the field of financial consumer behavior prediction.

3.2.1 Interpretability

Machine learning models are often likened to "black boxes" mainly because their decision-making process lacks transparency. In the financial field, the interpretability of the model is particularly important because it is related to key areas such as investment decisions and credit assessments. Financial institutions need to be able to explain the basis of their decisions to regulators and customers. For example, if a machine learning model is used for credit assessment and rejects a loan application, the financial institution needs to be able to explain the reasons behind this decision to ensure compliance and fairness. However, complex models often have difficulty providing this transparency, which limits their application in the financial sector.

3.2.2 Applicability

The quality and distribution of training data play a decisive role in the machine learning models. In the prediction of financial consumer behavior, if the training data does not fully represent all consumer groups, the model may produce biased prediction results. If the training data mainly comes from urban consumers, the model may not accurately predict the behavior of rural consumers. This distribution difference may lead to unfair financial services and limit the general applicability of the model.

3.2.3 Privacy

In the prediction of financial consumer behavior, machine learning models need to process a large amount of personal sensitive data, such as transaction records, personal identity information, etc. Due to the gradually increasing awareness of data privacy protection, finding the method to improve the security and privacy performance of this data has become a major challenge. Financial institutions ought to comply with strict data protection

regulations, such as the EU's General Data Protection Regulation (GDPR), which imposes requirements on the processing of data. In addition, with the development of technology, malicious users may also use machine learning to carry out more complex attacks. Therefore, it is essential to continuously update and strengthen the security of the model.

3.3 Future prospects

In summary, although machine learning provides powerful data analysis capabilities in the field of financial consumer behavior prediction, its limitations cannot be ignored. Financial institutions need to weigh these factors when applying these models and take corresponding measures in future research to ensure the fairness, compliance and security of the models. For example, future expert systems can further integrate machine learning techniques to automate the knowledge acquisition process and enhance the ability to handle complex problems. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are tools to improve the interpretability of models. SHAP assigns contributions to each feature through game theory principles, while LIME explains individual predictions through local linear models. In future research, it is possible to explore combining SHAP and LIME with other explanation methods to obtain more comprehensive model insights and develop real-time explanation engines to support the real-time explanation needs of high-volume applications. In the financial field, transfer learning methods are particularly worth learning, which allows the model to apply knowledge learned in one field to another different field, thereby helping the model predict effectively when data is scarce. Future research can explore more efficient transfer learning strategies to greatly enhance the accuracy and adaptability of the model in the target field. In general, the application prospects of machine learning in the field of financial consumer behavior prediction are optimistic. By combining the latest technological advances and innovative methods, existing challenges can be overcome.

4 Conclusion

This review explores the complex applications of machine learning in predicting financial consumer behavior, highlighting its transformative impact on the financial industry. In the process of customer churn prediction and loan prediction, models such as XGBoost and LightGBM have demonstrated excellent ability to handle complex data sets, and neural networks and ensemble methods such as random forests are becoming powerful tools. These models not only provide higher accuracy but are also able to handle imbalanced data and capture complex patterns, thereby enhancing risk assessment and credit scoring mechanisms. Through discussion, the study also analyzes limitations and challenges. The interpretability of some models, data distribution bias issues, and strict privacy issues deserve attention. Looking forward, the integration of machine learning with expert systems, the development of real-time explanation engines using SHAP and LIME, and the exploration of transfer learning strategies provide promising avenues for future research. This paper provides a good reference for the selection of prediction methods for machine learning models, which can facilitate relevant readers in the financial field to clarify future research directions. However, the article also has shortcomings. The current focus on deep learning models is relatively small, and some scenarios and models are not involved. Further, the research will be refined to form a more complete system for reference.

References

1. S. Khodabandehlou, M. Zivari Rahman, Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), 65-93 (2017).
2. C. Yu-Xuan, Prediction of Financial Customer Purchase Behavior Based on Machine Learning (in Chinese). Journal of Heihe University, 14(10), 52-56 (2023).
3. B. Gao, V. Balyan, Construction of a financial default risk prediction model based on the LightGBM algorithm. Journal of Intelligent Systems, 31(1), 767-779 (2022).
4. N. Romanenko, K. Sharma, S. Verma, Prediction of financial customer buying behavior based on machine learning. Journal of Artificial Intelligence General Science (JAIGS), 5(1), 125-131 (2024).
5. Z. Qing, Improvement of the Stacking Ensemble Algorithm Based on Generalization Accuracy (in Chinese). Xihua University, (2023).
6. O. F. Seymen, E. Ölmez, O. Doğan, et al., Customer churn prediction using ordinary artificial neural network and convolutional neural network algorithms: A comparative performance assessment. Gazi University Journal of Science, (2023).
7. S. K. Wagh, A. A. Andhale, K. S. Wagh, et al., Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342 (2024).
8. T. Gattermann-Itschert, U. W. Thonemann, Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests. Industrial Marketing Management, 107, 134-147 (2022).
9. K. Sravani, Using Random Forest as a Novel Approach to Loan Prediction and Comparing Accuracy to the Support Vector Machine Algorithm. Journal of Survey in Fisheries Sciences, 10(1S), 1174-1181 (2023).
10. M. Madaan, A. Kumar, C. Keshri, et al., Loan default prediction using decision trees and random forest: A comparative study. IOP Conference Series: Materials Science and Engineering, IOP Publishing, 1022(1), 012042 (2021).
11. R. Abbasov, Revolutionizing risk management in banking: Implementation of AI/ML-based gradient boosting machines (GBM) and random forest models for credit risk management, International Journal of Research in Finance and Management (2023)
12. M. Anand, A. Velu, P. Whig, Prediction of loan behaviour with machine learning models for secure banking. Journal of Computer Science and Engineering (JCSE), 3(1), 1-13 (2022).