Machine learning insights into digital payment behaviors and fraud prediction

Xu Zhang

Camegie Mellon University, Pittsburgh, Pennsylvania, United States

xuz5013@gmail.com

Abstract. With the continuous advancement of digital transformation, digital payments are playing an increasingly important role in the financial industry. This study aims to utilize machine learning models to predict and analyze digital payment behavior. Initially, the background and significance of digital payments in the financial sector are introduced. Subsequently, the current status and trends of traditional digital payment distribution are reviewed, alongside related work on digital payment behavior prediction. Methodologically, principles and applications of machine learning models such as logistic regression, decision trees, and random forests are elaborated, along with experimental design and data preprocessing methods. The experimental results and discussion section illustrates the performance of each model in digital payment prediction and explores their impact on credit decisions. This exploration equips financial institutions with more effective user behavior analysis and risk management tools, thereby fostering future development and application of digital payment technologies.

Keywords: Digital payments, machine learning models, forecasting, financial services.

1. Introduction

In a society with growing digital transformation, technology has demonstrated great potential in transforming money, payments and finance. As key stakeholders work to build public infrastructure to accelerate the transition to the digital economy, it is important to look inward and leverage existing development of technology to minimize adverse financial and economic impacts and contribute to the sustainable development of the society.[1]Digital payment breaks the time and space restrictions of traditional payment methods, making payment more convenient and efficient. Consumers simply scan a QR code or enter payment information via a mobile phone or other digital device to complete the payment. Compared with traditional cash payment and credit card payment, digital payment does not require change, signature or password input, which greatly shortens the payment process and improves payment efficiency.

For example, when shopping in a supermarket, consumers can use their mobile phones to scan the QR code on the product and complete the payment without waiting in line for the checkout. When dining at a restaurant, consumers can also use their mobile phones to scan a QR code on the table and complete the payment without waiting for a waiter.[2]With the continuous development of mobile Internet and intelligent technology, digital payment will show a broader development prospect and application scenario in the future. Enterprises should actively grasp the opportunities of digital payment, increase

digital transformation and innovation, improve payment services and user experience, and achieve sustainable business growth.

2. Related work

2.1. Traditional digital payment distribution

Digital payment is a payment method based on digital technology, which completes the transmission and processing of payment instructions through electronic means, so as to realize the transfer of funds. With the advantages of convenience, speed, efficiency and security, digital payment has become an important part of the modern payment system. [3]The current realizations of digital payment include electronic bank transfer, third-party payment, mobile payment and so on. Among them, the third-party payment is a relatively common digital payment method, which completes the liquidation and settlement of funds through the third-party payment platform. Digital payment is generally considered within the framework of electronic payment. While it differs from traditional forms of physical payment involving paper money or bills, there are further distinctions within digital payment itself.

This paper asserts that digital payment encompasses the application of various digital technologies, including automatic identification technology, information and communication technology, blockchain, big data, and cloud computing, within the payment domain. China's payment system comprises various elements such as payment tools, payment service providers, payment systems, payment supervision and management, as well as payment laws and regulations. Therefore, this paper focuses on the narrower interpretation of digital payment, emphasizing digital payment methods.blockchain, big data, and cloud computing in digital payment, whereas electronic payment emphasizes the issuance of payment orders via electronic terminals.

2.2. Trends in digital payments

The impetus behind the digital transformation of payment business primarily stems from several factors. Firstly, the global landscape is witnessing a profound industrial shift, propelled by the rapid advancement of digital technologies such as artificial intelligence, big data, cloud computing, blockchain, and the Internet of Things[4].As such, practitioners are progressively constructing convenient, efficient, and secure payment infrastructures, offering cost-effective payment tools, and fostering the integration of technology with the real economy along the trajectory of "technology-finance-industry"[5] development. Furthermore, they continually endeavor to build and refine the technology and financial ecosystem, with a focus on bolstering domestic demand, serving the tangible industries, and better catering to the personalized and diversified payment needs of individuals. Through these efforts, practitioners aim to deliver superior digital services and products, thereby enriching users' digital experiences.

1. Head Institutions: Head institutions maintain their leading position in user scale due to advantages in product matrix, closed-loop business models, technological innovation, scene integration, and channel sales. This has led to increasing concentration of user information. Moreover, the implementation of relevant laws and regulations, such as the Personal Information Protection Law and the Data Security Law, imposes stricter requirements and restrictions on user data collection, sharing, content push, and marketing promotion.

2. Enterprise Side: With the rapid advancement of online, digital, and intelligent payment businesses, practitioners are expanding their business horizons and seeking growth opportunities in traditional industries such as aviation, insurance, and retail. China's industrial payment market is projected to exceed 300 billion yuan in 2021, indicating a 40% year-on-year increase[6].

3. Merchant Side: Merchants are transitioning to multi-channel customer sales modes and integrated online and offline operations, leading to increased demand for public and private business linkage. Additionally, practitioners are enhancing merchant services from basic functions like account checking, QR code scanning, and ordering to value-added services aimed at revenue generation, such as technology, logistics, procurement, and finance support.

2.3. The challenges of digital payments

In the field of digital payments, practitioners face challenges in the accumulation, processing, analysis and sharing of data resources. Despite the huge potential of data in the current business environment, the popularity of data capitalization still needs to be strengthened. The norms and practices of ownership division, circulation, trading and sharing are still in the exploratory stage. At the same time, the improvement of data governance system and the strengthening of data security and privacy protection are also important challenges facing the current digital payment industry[7].

In addition to the challenges in terms of data resources, the digital payment industry is also facing challenges in terms of technology development.Therefore, one of the challenges in the future is to continue to explore new architectures and strengthen research on computing technology to meet the rapidly evolving needs of the digital payment industry. At the same time, the use of machine learning and other technologies to predict payment behavior and optimize the intelligence level of payment systems are also one of the important strategies to deal with technical challenges.

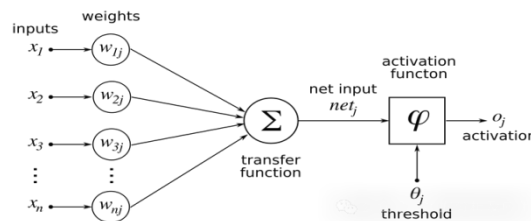## 3. Application of machine learning model in digital payment prediction

Machine learning is one of the many ways humans can achieve "artificial intelligence", trying to find a way to "predict" unknown data through empirical data. Traditional machine learning requires multi-domain knowledge and chooses appropriate methods such as decision tree, random forest and Bayesian learning according to the characteristics of data. Deep learning, on the other hand, is one of many methods of machine learning inspired by the way the human brain works.

3.1. Machine learning predictive models

In the realm of machine learning, diverse models, such as logistic regression, decision trees, support vector machines, etc., exhibit distinct applicability to various data types and problem types. Logistic regression excels in binary classification tasks, decision trees are proficient in handling both classification and regression tasks, while support vector machines demonstrate efficacy in addressing linear and nonlinear problems. Therefore, comprehending both the unique advantages and applicability scope of each model, as well as the commonalities and interrelationships between them, is pivotal for effective data analysis and prediction endeavors. Furthermore, in the context of predicting digital payment transaction fraud, considerations of model suitability, performance, and interpretability are paramount, with ensemble methods and deep learning models offering promising avenues for addressing the complexities inherent in such tasks.

(1) Neural network model

Neural network is a computational model that simulates the structure and function of human brain neural network. The basic unit is the neuron, and each neuron receives input from other neurons, changing the effect of the input on the neuron by adjusting the weight. The neural network is a black box, which can achieve the universal approximation effect through the action of multiple nonlinear hidden layers.

Figure 1. Architecture diagram of machine learning neural network model

Common machine learning models used in digital payment prediction models include deep neural networks (DNN)[8], support vector machines [9](SVM), Transformer, and long short-term memory networks (LSTM)[10]. DNNS are widely used to handle complex non-linear relationships and large-scale data sets, and their final layer is sometimes viewed as a logistic regression model for classifying

input data. Support vector machines can be regarded as a special type of neural network, which implements complex nonlinear transformations through kernel functions to achieve effects similar to deep neural networks.

In the realm of digital payment data analysis and prediction, Deep Neural Network (DNN) emerges as a widely utilized model comprising multiple layers of neurons. Through forward propagation, the DNN sequentially transmits input data to each layer of neurons and computes the output layer by layer. This facilitates the updating of neuron weights and bias terms according to the gradient descent algorithm, thereby iteratively optimizing network parameters to minimize prediction errors.

In the analysis and prediction of digital payment data, DNN offers several advantages: robust feature learning capability enables automatic extraction of data features without necessitating manual feature design; its highly nonlinear nature equips it with powerful generalization ability. However, DNN also presents some drawbacks. Despite these limitations, DNN's adeptness at automatically learning data features and modeling nonlinear relationships facilitates the effective capture of underlying rules in data, endowing it with potent modeling capabilities for predictive tasks, including the prediction of digital payment transaction fraud and other behaviors.

This ensemble approach enables the random forest to effectively handle complex nonlinear relationships and large-scale datasets, while demonstrating good generalization and anti-overfitting capabilities. Furthermore, compared to a single decision tree, the random forest model exhibits greater resilience and diminished susceptibility to outliers and noisy data. Consequently, the random forest model yields high prediction accuracy and reliability in predicting digital payment behavior, rendering it suitable for addressing intricate scenarios in practical applications, including the prediction of digital payment transactions and credit card fraud detection.

4. Methodology

In the field of financial digital payments, accurate prediction of user payment behavior is crucial to help financial institutions better understand user needs, manage risks and optimize services.   This study aims to explore and compare the performance of three commonly used machine learning models, XGBoost, decision tree and random forest, on financial digital payment datasets. Decision tree is a kind of nonlinear model based on tree structure, which can make hierarchical decision according to the payment characteristics of users.Random forest is an ensemble learning algorithm, which can combine the prediction results of multiple decision trees to improve the accuracy and robustness of the model. By comparing the classification prediction performance of these three models on financial digital payment datasets, we aim to identify the most suitable model for digital payment behavior prediction.

4.1. Experimental design

This experiment aims to evaluate the performance of two commonly used machine learning models, decision tree and random forest, by predicting fraudulent credit card transactions in digital payments. Fraudulent credit card transactions are one of the serious challenges facing the digital payment sector, causing serious losses and inconvenience to both financial institutions and users. By building predictive models, we hope to be able to identify suspicious transactions in a timely manner, thereby effectively preventing fraudulent events and improving the security and trust of digital payment systems.

This experiment will use real data sets from the financial digital payment field to compare and analyze the performance of decision tree and random forest model to determine the most suitable model for the prediction of credit card fraudulent transactions, so as to provide financial institutions with more reliable risk management tools and user protection measures.

4.2. DS

The data set is derived from real banking transactions of European cardholders in 2013. For security reasons, the original data has been transformed by PCA (Principal component analysis), and the transformed data set contains 29 feature columns and 1 class column. PCA transformation is a commonly used data dimensionality reduction technique, which converts raw data into a set of linearly

independent principal components through linear transformation to protect the privacy of users. These feature columns may cover a variety of factors related to the transaction, including the amount of the transaction, the time of the transaction, the location of the transaction, and so on.

4.3. Data processing

Data Characteristics and analysis: The data set is unbalanced because the majority of normal transactions in the data set are normal, and only a small percentage of transactions are fraudulent

Table1. Data distribution

Total number of Trnsactions are	84807
Number of Normal Transactions are	284315
Number of fraudulent Transactions are	492
Percentage of fraud Transactions is	0.17

In the data preprocessing stage, we first perform a basic information check on the data set and confirm whether there are missing values in the data through the 'data.info()' function. Based on the count of each column, we find no null values. We then further examined the feature columns of the dataset and found that 28 of the features were transformed by PCA, while the field "Amount" was raw. To eliminate the bias that this difference might have on the results, we normalized the "Amount" column by converting its value to a standard normal distribution with zero mean and unit variance.

4.4. Model building

We checked the data set for duplicate data and found that there were about 9,000 duplicate transactions. In order to avoid the influence of duplicate data on model building, we de-processed the data set. After de-duplication, our dataset contains cleaner transaction data, totaling 284,807 rows.

After data preprocessing is completed, we divide the data set into training set and test set. We define the independent variable as X, that is, remove all feature columns of the dependent variable "Class"; The dependent variable is defined as y, which is the "Class" column. We then use the 'train_test_split' function to split the data set into a training set and a test set at a ratio of 75% for training and evaluation of the model.

Next, we tried three different machine learning models: decision trees, random forests, and XGBoost. We define and train these models using the 'DecisionTreeClassifier', 'RandomForestClassifier', and 'XGBClassifier' classes, and evaluate their performance using test sets. For each model, we calculated Accuracy and an F1 score to assess the model's performance in predicting fraudulent credit card transactions. By comparing the performance metrics of these three models, we aim to identify the most appropriate model for predicting fraudulent credit card transactions to provide financial institutions with more reliable risk management tools and user protection measures.

4.5. Experimental discussion

Through The experiments, we have obtained a result of 99.95% accuracy for the credit card fraud detection model. However, this high accuracy rate is not surprising, as our dataset was designed for binary classification between normal and fraudulent transactions. In this scenario, even if the model predicts all transactions as normal, it can still achieve up to 99.83% accuracy. Therefore, accuracy alone may not be sufficient for evaluating the model's performance.

Based on our F1-Score analysis, the XGBoost model demonstrated superior performance in this context. However, it is essential to acknowledge that our dataset features were derived from PCA transformation. Consequently, our features may have lost some information from the original data and may no longer represent a linear combination of the original features. Thus, when assessing model performance, careful consideration of data characteristics and the impact of feature engineering is necessary. Further research and experimentation are warranted to explore more effective feature

selection methods and model tuning strategies, aiming to enhance the model's performance and robustness.

5. Conclusion

In summary, this study highlights the significant role of machine learning models in predicting and analyzing digital payment behavior within the financial industry. Through the application of various machine learning techniques such as logistic regression, decision trees, and random forests, the study demonstrates the efficacy of these models in improving the efficiency and accuracy of digital payment prediction. The experimental results underscore the importance of machine learning in enhancing user behavior analysis and credit decision-making processes, providing valuable insights for financial institutions seeking to optimize their services and manage risks effectively in the digital payments space.By leveraging blockchain technology, for example, financial institutions can enhance transaction transparency, traceability, and security, thereby mitigating fraud and enhancing user trust in digital payment platforms. Similarly, the utilization of big data analytics enables financial institutions to extract valuable insights from vast amounts of transaction data, facilitating more informed decision-making and personalized customer experiences in the realm of digital payments.

Furthermore, the future of digital payment prediction is closely intertwined with the broader trends shaping the financial industry, including the rise of digital currencies, the proliferation of IoT devices, and the increasing demand for seamless and frictionless payment experiences. As digital payment technologies continue to evolve and converge with other disruptive forces in the financial ecosystem, we can expect to see transformative changes in how payments are processed, authenticated, and secured in the years to come. By staying at the forefront of technological innovation and embracing new paradigms in machine learning and digital payments, financial institutions can position themselves for success in an increasingly digital and interconnected world.

References

[1] Patil, Pushp P., Yogesh K. Dwivedi, and Nripendra P. Rana. "Digital payments adoption: an analysis of literature." Digital Nations–Smart Cities, Innovation, and Sustainability: 16th IFIP WG 6.11 Conference on E-Business, E-Services, and E-Society, I3E 2017, Delhi, India, November 21–23, 2017, Proceedings 16. Springer International Publishing, 2017.

[2] Chaudhary, K., Yadav, J., & Mallick, B. (2012). A review of fraud detection techniques: Credit card. International Journal of Computer Applications, 45(1), 39-44.

[3] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (ICCNI)(pp. 1-9). IEEE.

[4] Carbo-Valverde, Santiago, Pedro Cuadros-Solas, and Francisco Rodríguez-Fernández. "A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests." Plos one 15.10 (2020): e0240362.

[5] Alexander, C., & Dakos, M. (2020). A critical investigation of cryptocurrency data and analysis. Quantitative Finance, 20(2), 173-188.

[6] Ma, Haowei. "Research on promotion of lower limb movement function recovery after stroke by using lower limb rehabilitation robot in combination with constant velocity muscle strength training." 2021 7th international symposium on mechatronics and industrial informatics (ISMII). IEEE, 2021.

[7] Sun, Y., Cui, Y., Hu, J., & Jia, W. (2018). Relation classification using coarse and fine-grained networks with SDP supervised key words selection. In Knowledge Science, Engineering and Management: 11th International Conference, KSEM 2018, Changchun, China, August 17–19, 2018, Proceedings, Part I 11 (pp. 514-522). Springer International Publishing.

[8] Huo, Mingda, et al. "JPX Tokyo Stock Exchange Prediction with LightGBM." Proceedings of the 2nd International Conference on Bigdata Blockchain and Economy Management, ICBBEM 2023, May 19–21, 2023, Hangzhou, China. 2023.

[9]   Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." The American Statistician 72.1 (2018): 37-45.
[10]   Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. IEEE Transactions on dependable and secure computing, 5(1), 37-48.
[11]   Patidar, R., & Sharma, L. (2011). Credit card fraud detection using neural network. International Journal of Soft Computing and Engineering (IJSCE), 1(32-38).