DOI: 10.53469/jtpes.2024.04(05).10

Utilizing Data Science and AI for Customer Churn Prediction in Marketing

Ang Li[1,*], Tianyi Yang[2], Xiaoan Zhan[3], Yadong Shi[4], Huixiang Li[5]

[1]Business Analytics, University College Dublin, Dublin, Ireland
[2]Financial Risk Management,University of Connecticut,stamford CT, USA
[3]Electrical Engineering, New York University, NY, USA
[4] Computer Science, Fudan University, ShangHai, China
[5]Information Studies, Trine University, AZ, USA
**Corresponding author, lalala19940211@gmail.com*

Abstract: ***This study explores the application of data science and AI techniques in predicting customer churn within the telecommunications industry, a sector characterized by intense competition and high customer turnover rates. By analyzing historical customer data, including usage patterns and service preferences, the study aims to identify factors contributing to churn and propose targeted retention strategies to mitigate losses. Traditional classification algorithms and ensemble techniques are evaluated using the Telecom-Customer-Churn dataset, with emphasis on the underutilized Stacking ensemble method. The results demonstrate that ensemble learning algorithms, particularly the Stacking model, outperform single algorithms, with CatBoost exhibiting the highest accuracy at 0.8119, followed closely by RandomForest at 0.7902 and XGBoost at 0.7820. These findings underscore CatBoost's superior generalization capabilities, likely attributed to its adept handling of categorical features and missing values, and its ability to model complex data relationships. The study contributes to advancing understanding of ensemble models and offers valuable insights for predicting telecom customer churn, thereby aiding in the development of effective retention strategies and enhancing customer satisfaction and loyalty.***

Keywords: Customer churn prediction, telecommunications industry, ensemble learning, CatBoost.

1. INTRODUCTION

In today's digital era, the telecommunications industry is highly competitive and saturated. Customers frequently switch providers to reduce costs or seek better services. By analyzing historical customer data, such as usage patterns and service preferences, companies can predict potential churn and implement targeted retention strategies, thus preventing unnecessary losses.[1-3]Data mining has seen rapid advancements and is widely used across various fields like finance and e-commerce. Applying these techniques to telecom customer churn can uncover the reasons behind churn and provide actionable insights for customer management and service improvement.Traditional classification algorithms such as Logistic Regression, Decision Trees, Random Forests, and Naive Bayes are often used for churn prediction. However, single algorithms may not achieve high accuracy. The Stacking ensemble technique, which combines multiple algorithms, can improve prediction performance and is underutilized in telecom churn prediction.

This paper conducts an in-depth analysis of the Telecom-Customer-Churn dataset, examining the relationship between variables and churn, and proposes strategies for churn prevention. Data preprocessing includes variable selection and balancing using [4]SMOTE. Various models are tested, including single algorithms (Logistic Regression, Decision Trees), and ensemble methods (LightGBM, Random Forests). For the Stacking model, Decision Trees, Random Forests, and [5]LightGBM serve as base classifiers, with Logistic Regression as the secondary classifier.

The results show that ensemble learning algorithms outperform single models, with the Stacking model providing the best performance. This research enhances the understanding of ensemble models and offers new insights for predicting telecom customer churn.

Driven by digital technology, Artificial Intelligence has ushered in rapid development, among which Generating

2. RELATED WORK

2.1 Customer churn forecasting

The full cost of churn includes revenue lost from old churn and marketing costs involved in replacing those customers with new ones, reducing churn is a key business objective for every company, and anticipating and preventing churn is a huge potential revenue source for every product/platform. [6]Customer loss analysis is mainly through analyzing user characteristics, looking for user characteristics that have a greater impact on user loss, and putting forward product/platform operation suggestions based on business knowledge in the field of e-commerce, so as to improve user stickiness and reduce user loss rate.

Churn analysis is an ongoing effort that requires [7]long-term monitoring and iteration. It is necessary to regularly monitor customer behavior and abnormal indicators, and adjust user policies based on data feedback. The business environment and market needs are constantly changing, including other products in the industry, so the analysis methods and conclusions are constantly changing.

Also, be aware that there is some lag in the data. From the adjustment of products and services, to users receiving feedback, to data collection and indicator changes, the whole process takes a certain amount of time. [8]Therefore, in addition to relying on numbers, it is also necessary to have business foresight and sensitivity to understand user needs and situations in order to judge and predict before data.

Step 1: Data collection

Collect relevant data: In this step, we need to collect data related to customer interactions, transactions and behavior. This includes purchase history, usage patterns, customer interactions, and demographic information.
Data sources: Data can come from a variety of sources, we can use existing customer relationship management systems, transaction logs, customer research and other channels. The integration of these data sources will provide us with a more comprehensive and accurate view of the data.

Step 2: Define churn

Define churn metrics: [9\Clearly define churn, such as not making a purchase for a period of time, canceling a subscription, or expressing unsatisfactory feedback.

Churn time frame: Determine a time frame to measure churn, such as monthly, quarterly, or yearly, depending on the characteristics and needs of the business.

Step 3: Data cleaning and preprocessing

This includes processing missing data, removing duplicates, removing outliers, and other pre-processing operations to identify and remove outliers that may distort the analysis results to ensure the accuracy of the analysis.

Step 4: Feature selection

Identify relevant characteristics: [10]Identify key data characteristics relevant to churn analysis, such as usage frequency, purchase history, customer demographic information, and customer service. Remove some minor, irrelevant data characteristics, such as user ID, user name, etc.

Correlation analysis: Correlations between features are analyzed to determine the extent to which they contribute to churn.

Step 5: Exploratory Data Analysis (EDA)[11]

Visualize data: Explore patterns and trends in customer behavior using data visualization techniques such as histograms and scatter plots.

Descriptive statistics: [12] Calculate descriptive statistics for key variables to understand their distribution and central trends.

Step 6: Build a predictive model

Training/test set segmentation: Data is divided into training sets and test sets for training and evaluating the performance of predictive models.

Choose a model: Choose a predictive model suitable for churn analysis, such as logistic regression, decision trees, or machine learning algorithms.

Feature importance: [13]Analyze the significance of features to understand which factors are most critical for predicting churn.

Step 7: Model evaluation

Metrics selection: The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1 scores, and adjusted as needed.

Step 8: Interpret the results

Identify churn factors: By interpreting the output of the model, we can determine the key factors that affect churn and the extent to which they affect it. This helps us understand the reasons for customer churn and develop strategies to respond accordingly.

Step 9: Implement mitigation strategies

Develop retention strategies: [14]One of the goals of churn analysis is to develop a targeted retention strategy based on the results of the analysis. This may include personalized services, membership programs, product service improvements and other measures aimed at increasing customer satisfaction and reducing attrition rates.

Step 10: Monitor and iterate

Continuous monitoring: Regularly monitor churn metrics and customer behavior and adjust retention strategies based on ongoing data analysis.

Iterative analysis: The churn analysis process is constantly iterated to adapt to the arrival of new data or changes in business conditions. [15]This means that we need to constantly learn and improve to build smarter and more adaptable attrition analysis systems.

2.2 Differences between churn analysis and other analyses

When conducting churn analysis, it differs from other user or sales analysis in its focus and perspective.

1. Differences in emphasis:

Sales data and overall user data analytics typically focus on understanding current sales trends, customer behavior, and overall market performance. [16]These analyses are mainly used to evaluate performance and market share, and help companies develop marketing strategies and sales plans.

Churn analysis is more focused on exploring the causes and patterns of churn. It is a problem-oriented approach to analysis that focuses on why customers choose to leave and how to prevent or reduce this churn. [17]Churn analysis aims to identify potential trouble spots and take action to retain existing customers.

2. Different analysis angles:

Product sales analysis is from the perspective of product and sales performance, focusing on product characteristics, market trends and competition. The lack of perspective from the customer's point of view. In the product dimension, we can find out how the product performs and whether the sales result is popular.

However, from an individual user's perspective, whether they want to continue transacting with us or using our services may be affected by a number of factors. [18]When deciding whether to quit a game or stop using a service,

it is usually not due to the poor performance of a single product, but may be a combination of factors. For example, the combination of poor after-sales service and problems with the product itself can lead to customer churn.

In addition, changes in the user's personal life and work may also cause them to no longer need a service, and such changes do not represent a problem with the quality of the product or service. In this case, optimizing their own products or services may not be the best way to solve the problem, but it is more important to understand the market needs and meet the market needs.

In summary, user churn analysis is from the customer's perspective, focusing on customer experience, satisfaction and loyalty. It focuses more on understanding customer needs, behavior and feedback to improve customer retention and loyalty. [19]Therefore, when conducting user analysis, we need to collect as much data as possible, which covers not only the product aspect, but also the personal information of the customer, the user experience, and so on. This data is a record of all interactions between users and businesses.

3. Characteristics and unique value of user churn analysis:

In-depth exploration of potential problems: User loss analysis by in-depth exploration of the causes and patterns of customer loss, help enterprises to find potential problems, so as to take timely measures to solve them.

Improve customer experience:[20] By understanding customer needs and behaviors, churn analysis can help optimize products and services, enhance customer experience and satisfaction, and thereby enhance customer loyalty and long-term value.

Reduce cost risk: Preventing customer churn is more cost-effective than attracting new customers. Churn analysis can help companies reduce churn rates and reduce marketing and customer acquisition costs.

Improve competitiveness: [21]By continuously improving products and services, as well as maintaining customer satisfaction and loyalty, companies can improve their position and competitiveness in a highly competitive market.

2.3 Customer churn warning model

The loss early warning model needs to adopt different models to forecast users with different life cycles, which can be divided into acquisition period, promotion period, maturity period and decline period. The purpose of the cycle division is to incorporate the user life stage into the refined operational early warning recall strategy in the future. Loss warning is to extract historical user data, observe the relevant data in a certain window, and then evaluate the loss of users in the performance window according to the above loss user definition, so as to predict the loss probability of current users in the future.

So what user data can affect user churn? It can be roughly divided into three dimensions, namely, user portrait data, user behavior data, and user consumption data. In addition, we need to define the forecast time window, i.e. how long period of time should we analyze the sample data? It is necessary to combine the experience of business personnel and historical user behavior data, and then synthesize the availability of data, and finally establish a reasonable time prediction window[22-24].

First of all, it is necessary to mine a group of sample users from historical data, and improve evaluation indicators at all levels according to the three main dimensions of user portrait data, user behavior data and user consumption data, so as to cover a full range of field data as far as possible, so as to facilitate the evaluation of the correlation between indicators and loss in subsequent modeling. [25]By obtaining the result data in the performance period window, the final prediction model can be built, and the user loss rule and the importance ranking of each feature index can be obtained. Common early warning algorithms include decision tree, random forest, logistic regression and so on. During the prediction window, we continuously optimize the trained model and eliminate some features with low correlation. The accuracy, hit rate and coverage rate of the model are improved, and then the probability of user loss in the next month can be predicted, and the score and list of lost users can be output.

3. CUSTOMER CHURN PREDICTION MODEL IN TELECOM INDUSTRY

In the telecommunications industry, customers have the flexibility to choose from various service providers. Customer churn is defined as the scenario where a customer ceases to engage in business with a company or

discontinues using its services. [26]The ability to predict customer churn is crucial for telecom companies as it allows them to implement retention strategies and reduce the impact of losing customers to competitors.

3.1 Experimental design

This project aims to develop a predictive model for customer churn using the provided dataset. By analyzing historical customer data, such as service usage patterns, billing information, and customer interactions, we can identify the factors that contribute to churn. The goal is to leverage this analysis to predict which customers are likely to churn in the future.

Approach involves several key steps[27]:

1. Data Preprocessing: Cleaning and preparing the data for analysis, including handling missing values, normalizing data, and encoding categorical variables.

2. Exploratory Data Analysis (EDA)[28]: Examining the relationships between various features and customer churn to identify significant predictors.

3. Feature Engineering: Creating new features based on existing data to improve the model's predictive power.

4. Model Building: Utilizing various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, and ensemble methods like Stacking, to build and compare predictive models.

5. Model Evaluation: Assessing the performance of the models using appropriate metrics such as accuracy, precision, recall, and F1-score.

By integrating data science and artificial intelligence techniques, we aim to build a robust model that can accurately predict customer churn, providing telecom companies with actionable insights to enhance customer retention strategies. The results of this project will not only help in understanding the factors leading to customer churn but also demonstrate the effectiveness of advanced modeling techniques in addressing this critical business challenge.

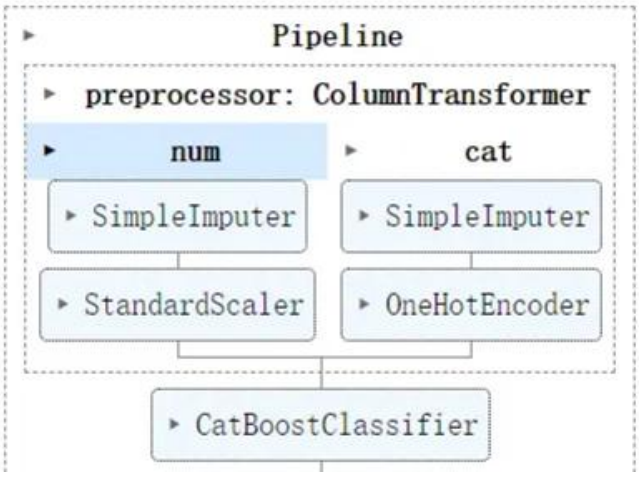

Figure 1. Flow chart of experimental steps

3.2 Experimental result

The results on the test set showed an accuracy of about 81%. After evaluation, we found that the CatBoost model performed best, with an accuracy of 0.8119. This shows that CatBoost has demonstrated quite strong predictive power on our dataset. This result is very encouraging for our mission and shows that we have chosen a valid model to solve the problem. CatBoost's excellent performance may be attributed to its ability to handle category features and missing values, as well as its ability to handle complex data relationships. This result provides a solid foundation for future work to further optimize the model or explore deeper data analysis.

Table 1. Data table of experimental results

Model	Accuracy
RandomForest	0.79
XGBoost	0.782
CatBoost	0.802

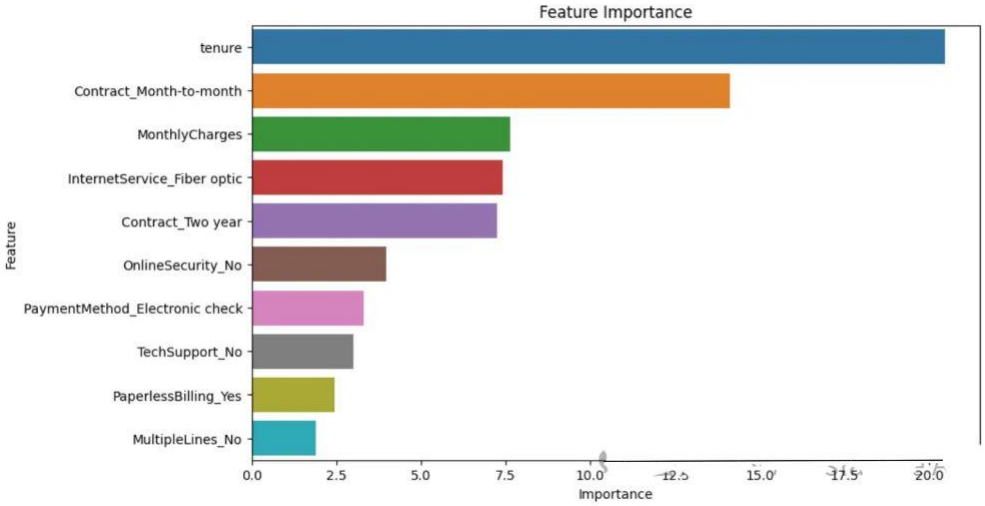

Figure 2. Model diagram of training results

Based on the results of our experiments with customer churn prediction, we evaluated three different machine learning models: RandomForest, XGBoost, and CatBoost. The experimental results show that the accuracy of these models on the test set is stable at about 80%, indicating that they are effective in predicting customer churn. Most encouragingly, however, the CatBoost model performed best, achieving an accuracy of 0.8119, significantly better than the other two models. The Random forest model is a close second with an accuracy of 0.7902, while XGBoost is slightly less accurate at 0.7820. These results show that CatBoost exhibits stronger generalization capabilities on our dataset, possibly due to its ability to handle class features and missing values, as well as its ability to model complex data relationships. These findings provide important clues for us to choose the right machine learning model and provide useful references for future experiments and applications in the field of customer churn prediction.

4. CONCLUSION

In this study, we explore the application of data science and artificial intelligence techniques to predict customer churn in the telecommunications industry. By analyzing historical customer data, including usage patterns and service preferences, we identify the factors that influence customer churn and propose targeted retention strategies to mitigate losses. [29-33]The experimental results show that integrated learning algorithms, especially the Stacking model, perform better than single algorithms. CatBoost ranks first with the highest accuracy of 0.8119, followed by RandomForest's 0.7902 and XGBoost's 0.7820. These findings highlight CatBoost's superior performance in handling categorical features and missing values, as well as modeling complex data relationships. The study provides valuable insights into improving the understanding of integrated models for telecom customer churn forecasting, thereby helping to develop effective retention strategies to improve customer satisfaction and loyalty.

In summary, this study provides a comprehensive and in-depth look at customer churn forecasting in the telecom industry, demonstrating the potential of data science and artificial intelligence in solving this critical business challenge. By adopting integrated learning algorithms such as CatBoost[34], we not only improve prediction accuracy, but also provide important guidance for future experiments and applications. Our research provides telecom companies with practical tools and methods to better understand and respond to customer churn, thereby boosting business growth and enhancing competitiveness.

ACKNOWLEDGEMENT

We would like to thank Penghao Liang, Bo Song, Xiaoan Zhan, Zhou Chen and Jiaqiang Yuan for their hard work and outstanding contributions. They are in [1]"Automating the Training and Deployment of Models in MLOps by Integrating Systems with Machine Learning" The key research results presented in this article have made important contributions to the development of the MLOps field.

Thank you for the inspiration and help of this article. Their research has provided us with valuable insights and concepts that help guide our work in the field of machine Learning Operations (MLOps). We are grateful for their outstanding work and look forward to working with them to further the development of this field in the future.

In addition, sincere thanks to Yanlin Zhou, Xinyu Shen, Kai Tan, Zheng He and Haotian Zheng wrote for [2]"A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration". The efforts and contributions made by this article. Their research results have brought important innovations to the field of protein structure prediction and have made important contributions to the progress of the scientific community.

At the same time, we also want to thank their research for our inspiration and help. Their work has provided us with valuable ideas and methods, and has played a guiding role in our research on protein structure prediction. We sincerely appreciate their outstanding contributions and look forward to exploring more possibilities in this field with them in the future."

REFERENCES

[1] Liang, P., Song, B., Zhan, X., Chen, Z., & Yuan, J. (2024). Automating the Training and Deployment of Models in MLOps by Integrating Systems with Machine Learning. arXiv preprint arXiv:2405.09819.
[2] Zhou, Y., Tan, K., Shen, X., & He, Z. (2024). A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration. arXiv preprint arXiv:2402.19095
[3] Lei, H., Chen, Z., Yang, P., Shui, Z., & Wang, B. (2024). Real-time Anomaly Target Detection and Recognition in Intelligent Surveillance Systems based on SLAM.
[4] Yang, P., Shui, Z., Chen, Z., Baoming, W., & Lei, H. (2024). Integrated Management of Potential Financial Risks Based on Data Warehouse. Journal of Economic Theory and Business Management, 1(2), 64-70.
[5] Shen, X., Wang, B., He, Z., Zhou, H., & Zhou, Y. (2024). Biology-based AI Predicts T-cell Receptor Antigen Binding Specificity. Academic Journal of Science and Technology, 10(1), 23-27.
[6] Lei, Han, et al. "Automated Lane Change Behavior Prediction and Environmental Perception Based on SLAM Technology." arXiv preprint arXiv:2404.04492 (2024).
[7] Shen, Xinyu, et al. "Biology-based AI Predicts T-cell Receptor Antigen Binding Specificity." Academic Journal of Science and Technology 10.1 (2024): 23-27.
[8] Lei, H., Wang, B., Shui, Z., Yang, P., & Liang, P. (2024). Automated Lane Change Behavior Prediction and Environmental Perception Based on SLAM Technology. arXiv preprint arXiv:2404.04492.
[9] He, Z., Shen, X., Zhou, Y., & Wang, Y. Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering.
[10] Sha, X. (2024). Time Series Stock Price Forecasting Based on Genetic Algorithm (GA)-Long Short-Term Memory Network (LSTM) Optimization. arXiv preprint arXiv:2405.03151.
[11] Cui, Z., Lin, L., Zong, Y., Chen, Y., & Wang, S. Precision Gene Editing Using Deep Learning: A Case Study of the CRISPR-Cas9 Editor.
[12] Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J., & Wan, W. (2024). Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management. arXiv preprint arXiv:2402.17216.
[13] Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B., & Zhang, Y. (2024). Dynamic Resource Allocation for Virtual Machine Migration Optimization using Machine Learning. arXiv preprint arXiv:2403.13619.
[14] Huang, J., Zhang, Y., Xu, J., Wu, B., Liu, B., & Gong, Y. Implementation of Seamless Assistance with Google Assistant Leveraging Cloud Computing.
[15] Zhou, Y., Zhan, T., Wu, Y., Song, B., & Shi, C. RNA Secondary Structure Prediction Using Transformer-Based Deep Learning Models.
[16] Liu, B., Cai, G., Ling, Z., Qian, J., & Zhang, Q. Precise Positioning and Prediction System for Autonomous Driving Based on Generative Artificial Intelligence.
[17] Wang, X., Tian, J., Qi, Y., Li, H., & Feng, Y. (2024). Short-Term Passenger Flow Prediction for Urban Rail Transit Based on Machine Learning. Journal of Computer Technology and Applied Mathematics, 1(1), 63-69.

[18] Zhou, Hong, et al. "Application of Conversational Intelligent Reporting System Based on Artificial Intelligence and Large Language Models." Journal of Theory and Practice of Engineering Science 4.03 (2024): 176-182.
[19] Xu, Kangming, et al. "Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning." arXiv preprint arXiv:2403.19345(2024).
[20] Lin, T., & Cao, J. * "Touch Interactive System Design with Intelligent Vase of Psychotherapy for Alzheimer's Disease," Designs, 2020, 4(3), 28. Journals Designs Volume 4 Issue 3 10.3390/designs4030028
[21] Li, H., Wang, X., Feng, Y., Qi, Y., & Tian, J. (2024). Driving Intelligent IoT Monitoring and Control through Cloud Computing and Machine Learning. arXiv preprint arXiv:2403.18100.
[22] Yu, D., Xie, Y., An, W., Li, Z., & Yao, Y. (2023, December). Joint Coordinate Regression and Association For Multi-Person Pose Estimation, A Pure Neural Network Approach. In Proceedings of the 5th ACM International Conference on Multimedia in Asia (pp. 1-8).
[23] Zheng, Haotian, et al. "Medication Recommendation System Based on Natural Language Processing for Patient Emotion Analysis." Academic Journal of Science and Technology 10.1 (2024): 62-68. （10个）
[24] Li, Lianwei, et al. "Independent Grouped Information Expert Model: A Personalized Recommendation Algorithm Based on Deep Learning." (2024).
[25] Ding, W., Zhou, H., Tan, H., Li, Z., & Fan, C. (2024). Automated Compatibility Testing Method for Distributed Software Systems in Cloud Computing.
[26] Qian, Kun, et al. "Implementation of Artificial Intelligence in Investment Decision-making in the Chinese A-share Market." Journal of Economic Theory and Business Management 1.2 (2024): 36-42.
[27] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing.
[28] Feng, Y., Li, H., Wang, X., Tian, J., & Qi, Y. (2024). Application of Machine Learning Decision Tree Algorithm Based on Big Data in Intelligent Procurement.
[29] Ni, C., Zhang, C., Lu, W., Wang, H., & Wu, J. (2024). Enabling Intelligent Decision Making and Optimization in Enterprises through Data Pipelines.
[30] Bao, Q., Wei, K., Xu, J., & Jiang, W. (2024). Application of Deep Learning in Financial Credit Card Fraud Detection. Journal of Economic Theory and Business Management, 1(2), 51-57.
[31] Fan, Chao, et al. "Integrating Artificial Intelligence with SLAM Technology for Robotic Navigation and Localization in Unknown Environments."
[32] Bai, X., Jiang, W., & Xu, J. (2024). Development Trends in AI-Based Financial Risk Monitoring Technologies. Journal of Economic Theory and Business Management, 1(2), 58-63.
[33] Tian, J., Qi, Y., Li, H., Feng, Y., & Wang, X. (2024). Deep Learning Algorithms Based on Computer Vision Technology and Large-Scale Image Data. Journal of Computer Technology and Applied Mathematics, 1(1), 109-115.
[34] Wang, Y., Zhu, M., Yuan, J., Wang, G., & Zhou, H. (2024). The intelligent prediction and assessment of financial information risk in the cloud computing model. arXiv preprint arXiv:2404.09322.