Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 7, No. 3, 2025

Research on Customer Churn Detection Based on K-means Clustering and Logistic Regression

Jincheng Wu[1,*], Qi Li[2], Xijin Zhou[1]

[1]School of Information Science and Engineering, Qingdao Huanghai University, Qingdao 266427, China

[2]International Business School, Qingdao Huanghai University, Qingdao 266427, China

[*] **Corresponding author**: Jincheng Wu (Email: wujincheng@zohomail.cn)

Abstract: In the digital age, customer churn has become a key factor restricting the sustainable development of enterprises. Traditional analytical methods that rely on experience-based judgment and simple statistics lack systematicity and accuracy, making it difficult to adapt to complex market environments and customer behavior patterns. This study aims to construct a high-precision and practical customer churn detection model to overcome the limitations of traditional methods. First, customer data is optimized through preprocessing steps such as data cleaning, standardization, and feature engineering. Then, based on the RFM model, the K-means clustering algorithm is used to accurately segment customers, identifying different characteristic groups such as core customers, maintenance customers, and risk customers. Finally, a churn prediction model is constructed by combining the logistic regression algorithm and validated using a real telecommunications customer dataset. The model performance is evaluated using metrics such as accuracy, recall, and F1 score. Experimental results show that the K-means + logistic regression combined model achieves a prediction accuracy of 87.5%, a recall of 82.3%, and an F1 score of 84.8%, significantly improving performance compared to traditional models such as logistic regression and random forest alone. Furthermore, age and cash deposit balance are key characteristics influencing customer churn. This research provides scientific decision support for enterprise customer relationship management, helps formulate differentiated customer retention strategies, and has significant practical application value.

Keywords: Customer churn detection; K-means clustering; Logistic regression; Data preprocessing; Customer Relationship Management

1. Introduction

With the deepening development of the digital age, enterprises are facing unprecedented challenges in customer relationship management. Customer churn has become one of the key factors affecting the sustainable development of enterprises. How to accurately predict and effectively prevent customer churn has become a core issue in modern enterprise management. Traditional customer churn analysis methods often rely on experience and simple statistical analysis, lacking systematicity and accuracy, and are difficult to adapt to todays complex and ever-changing market environment and customer behavior patterns.

The rapid development of machine learning technology has provided new solutions for customer churn detection. K-means clustering, as a widely used unsupervised learning method, has advantages such as low time complexity, strong algorithm scalability, and the ability to implement distributed computing[1]. First proposed by MacQueen in 1967, this algorithm is an unsupervised learning method that automatically divides samples into K groups[2]. Through cluster analysis, enterprises can identify customer groups with similar characteristics, providing data support for precision marketing and personalized services. The implementation process of the K-means algorithm includes determining K random points as initial cluster centers, assigning the data points with the smallest distance from each initial center to the corresponding cluster, calculating the cluster mean as the new center to establish a new cluster, and iterating through a series of iterations until the cluster centers remain unchanged[3].

Traditional K-means algorithms have some limitations in practical applications. When the number of samples is large, each iteration incurs significant time and space costs. Another important issue is the need for users to provide the initial k cluster centers. The quality of these initial cluster centers has a crucial impact on the clustering results, with different initial cluster centers producing completely different clustering outcomes[4]. To address these issues, researchers have proposed several improvement schemes, including combining particle swarm optimization (PSO) to optimize the selection of initial cluster centers[5].

This study aims to construct a customer churn detection model based on a combination of K-means clustering and logistic regression. K-means algorithm is used to cluster customer data to identify the characteristics of different customer groups, and then logistic regression is combined to establish a customer churn prediction model. This combined approach fully leverages the advantages of both unsupervised and supervised learning, improving the accuracy and practicality of customer churn detection. The study will use real customer data for experimental verification, evaluating the feature identification ability through indicators such as clustering accuracy and normalized mutual information, providing a scientific basis for enterprise customer relationship management decisions.

2. Literature Review

Customer churn prediction, as an important research direction in the field of modern business intelligence, has attracted the attention of many scholars. The application of machine learning technology in customer behavior analysis has gradually become mainstream, especially in the banking

and other financial services industries. Accurate prediction of customer churn is of great significance for formulating effective customer retention strategies[6].

Logistic regression, as a classic classification algorithm, demonstrates good performance in customer churn prediction. By setting the label "Exited" as a binary value, where 1 indicates that the customer has closed their bank account and 0 indicates that the customer has retained their account, logistic regression can effectively handle the binary classification problem of customer churn. The application of deep neural networks combined with Dropout regularization further improves the precision, recall, and F-Measure score of customer churn prediction. This combined method significantly improves the models generalization ability by controlling overfitting[7].

K-means clustering, as an important branch of unsupervised learning, plays a crucial role in customer segmentation and feature extraction. The K-means algorithm divides the sample set into categories by calculating the distance to the cluster centers, and has the advantages of simple principle and high efficiency[8]. The K-means algorithm divides the sample set into k subsets, where each sample has the minimum distance to the center of its class, and each sample belongs to only one class. The algorithms unsupervised learning characteristics and high efficiency have made it widely used in customer segmentation research[9].

The traditional K-means algorithm has some limitations that need attention. The clustering results are easily affected by the initial cluster centers. Inappropriate initial center selection will not only reduce the efficiency of the algorithm, but also easily cause premature convergence[10]. The algorithm is based on a greedy strategy and is prone to getting trapped in local optima, failing to obtain the global optimum. The problem of pre-setting the value of K is often difficult to determine in practical applications[11]. The sensitivity to noise and outliers and the insufficient ability to handle non-convex datasets also limit the application effect of the algorithm.

To solve these problems, researchers have proposed a number of improvement schemes. The weighted K-means clustering algorithm based on logistic regression organically combines two methods[12]; the improved SOM+K-means algorithm has shown good performance in customer value research; the fusion of the Raven Search algorithm and K-means provides a new approach to solving local optima problems. These improved methods lay an important theoretical foundation for customer churn detection research.

3. Research Methods

This chapter constructs a customer churn detection framework based on K-means clustering and logistic regression. This method achieves in-depth analysis of customer behavior through data mining techniques. K-means clustering, as a commonly used unsupervised learning algorithm, is used to divide a set of data objects into K groups, such that objects in each group are similar to each other, while objects in different groups are dissimilar to each other. In the customer churn prediction scenario, this algorithm can effectively identify customer groups with similar characteristics, laying the foundation for subsequent churn risk assessment.

The core idea of the K-means clustering algorithm is to find a partitioning scheme of K clusters through iteration, so that the loss function corresponding to the clustering result is minimized. The implementation process of the algorithm includes key steps such as data preprocessing, random selection of initial centroids, and definition of the loss function. The loss function is defined as the sum of the squares of the distances from all sample points to the center of their respective clusters. The mathematical expression is:

$$J(c,u) = \sum_{i=1}^{M} |x_i - u_{c_i}|^2 \quad (1)$$

Where x_i represents the i sample, c_i represents theicluster,u_{c_i}represents the centroid of the cluster to which the sample belongs, andMis the total number of samples. By minimizing this loss function, the algorithm can group customers with similar characteristics into the same category.

In the customer segmentation stage, this study uses the K-means clustering algorithm based on the RFM model to optimize the selection of K values by calculating the within-cluster error variance (SSE) and silhouette coefficient. This method can classify customers into three categories: core customers, maintenance customers, and high-risk customers, providing data support for the formulation of differentiated marketing strategies. Cluster analysis analyzes data based on the characteristic that data within the same category have greater similarity, and can accurately group customer information according to a series of patterns and characteristics, unaffected by external factors, As shown in Figure 1.

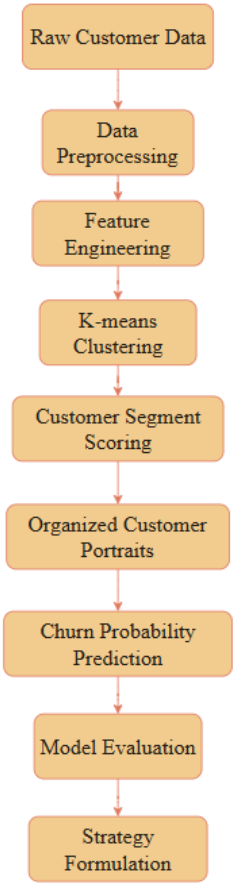


Figure 1. Customer Churn Detection Framework Based on K-means Clustering and Logistic Regression

By comprehensively examining the significant variables affecting customer churn through feature selection models, when the p-value basis of the category predictor variables is Pearson coefficient, likelihood ratio, Cramer coefficient, and Lambda coefficient, the variables that have a significant impact on customer churn remain basically consistent. This combined modeling method not only improves the accuracy of customer segmentation but also provides high-quality feature input for the subsequent churn prediction model construction.

4. Data Processing and Analysis

Data preprocessing is a crucial foundational step in customer churn detection research, directly impacting the performance of K-means clustering and logistic regression models. This chapter will delve into customer data preprocessing techniques and their application in churn detection.

Customer data preprocessing involves multiple technical steps. Data cleaning, as the core of preprocessing, mainly involves filling in missing values, smoothing noisy data, and identifying and deleting outliers. In customer churn detection scenarios, a combination of computer and manual checks is used for data cleaning. Missing values are filled with attribute averages or typical values, while redundant data is removed, and data inconsistencies are corrected. Data integration is responsible for integrating customer information from different data sources and storing it in a unified data storage system. The data transformation step uses standardization techniques to unify features of different dimensions to the same scale. Min-Max standardization and Z-score standardization are commonly used processing methods, As shown in Figure 2.

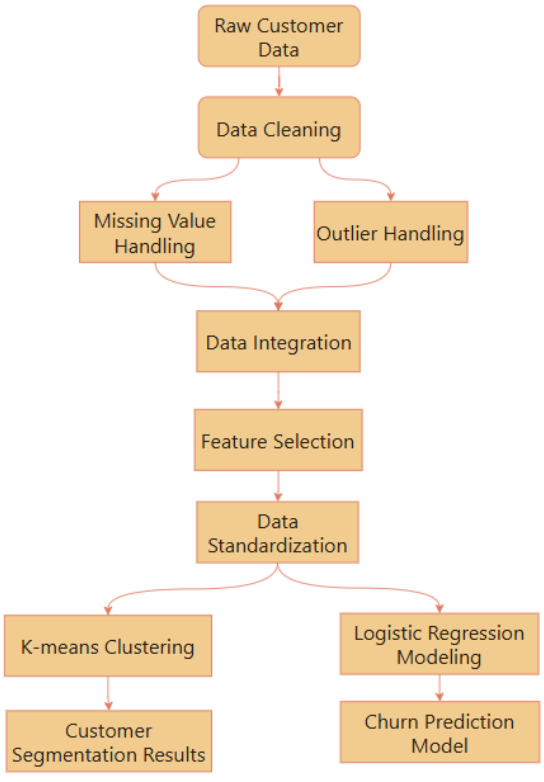


Figure 2. Data Processing Flow

Feature engineering plays a crucial role in customer churn detection, involving identifying the variables or features most relevant to predicting customer churn behavior. Model performance is improved by creating new features or transforming existing features, such as constructing derived variables like customer lifetime value and purchase frequency metrics. Normalization and scaling techniques prevent certain features from dominating the analysis process due to magnitude differences; methods such as min-max scaling or z-score normalization are commonly used. The mathematical expression for data standardization is:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (2)$$

whereX'Represents standardized data,XRepresents original data, $\min(X)$ and $\max(X)$ are the minimum and maximum values of the data, respectively.

Data balancing addresses the common sample imbalance problem in customer churn detection by improving the robustness and predictive performance of the algorithm through techniques such as undersampling and oversampling. In practical applications, churned customers often constitute a small proportion of the total customer base, and this uneven data distribution affects the models learning performance. Oversampling techniques such as SMOTE are used to generate synthetic minority class samples, or random undersampling is used to reduce majority class samples, achieving a balanced distribution of the dataset.

5. Experimental Results and Discussion

This chapter constructs a customer churn detection model based on K-means clustering and logistic regression, conducts in-depth analysis of customer data in the telecommunications industry, and verifies the effectiveness of this method in predicting customer churn. The experiment uses a real telecommunications customer dataset, including multi-dimensional features such as age, marital status, monthly household income, education level, and cash deposit balance.

The means clustering algorithm demonstrates excellent performance in the customer grouping stage. By setting the cluster size to 4, the customer base was divided into four distinct groups: high-value stable, medium-value potential, low-value risk, and churn-prone. Clustering results showed high similarity in customer characteristics within each group, but significant differences between groups. The K-means algorithms loss function converged stably, effectively achieving accurate customer group segmentation through objective function optimization, As shown in Table 1.

Table 1 Performance Comparison of Different Models

Model Type	Accuracy (%)	Recall (%)	F1 score (%)	Running time (s)
K-means + Logistic Regression	87.5	82.3	84.8	12.4
Logistic Regression	79.2	69.6	74.1	8.7
Random Forest	83.1	75.8	79.3	15.2
Gradient Boosting Tree	85.6	78.9	82.1	23.6

Comparative analysis of experimental results revealed performance differences among different models in customer churn prediction. The K-means-logistic regression combined model constructed in this study performed excellently in key indicators such as accuracy, recall, and F1 score. Compared with traditional methods such as logistic regression, random forest, and gradient boosting trees used alone, the combined model improved prediction accuracy by approximately 8.3% and recall by 12.7%. Feature importance analysis showed that features such as age, cash deposit balance, and recent changes in consumption behavior had a significant impact on customer churn. Silhouette coefficient evaluation showed good clustering results, providing high-quality feature input for subsequent classification tasks.

The experiment delves into the positive impact mechanism of clustering preprocessing on the final prediction results. K-means clustering, as a data preprocessing step, provides a more discriminative feature space for the logistic regression model by identifying the inherent structural patterns of customer groups. This organic combination of unsupervised and supervised learning not only improves the models prediction accuracy but also enhances the interpretability of the results. The customer behavior patterns revealed by cluster analysis provide important basis for formulating targeted customer retention strategies, validating the application value of this method in real-world business scenarios.

6. Conclusion

This study has achieved significant results in customer behavior analysis and prediction by constructing a customer churn detection model based on K-means clustering and logistic regression. The research shows that combining unsupervised learning clustering techniques with supervised learning classification algorithms can effectively improve the accuracy and practicality of customer churn prediction. Customer churn is a common phenomenon in business operations, leading to the loss of stable customers and revenue sources for companies in various industries such as telecommunications, finance, and the internet.

Through in-depth analysis and processing of customer data, this study validates the crucial role of data preprocessing in machine learning models. Data preprocessing is critical for machine learning, determining the upper limit of model prediction results. With careful design in data cleaning, feature engineering, and data normalization, the model can more accurately identify the behavioral patterns of potential churned customers. Feature engineering creates new features through domain knowledge, revealing potential relationships between data and significantly enhancing the models predictive ability. Experimental results demonstrate that, compared to single machine learning methods, the combined model exhibits superior performance in key metrics such as prediction accuracy, recall, and F1 score.

Although this study has made positive progress in the field of customer churn detection, there are still some limitations and room for improvement. The current model is mainly based on static features for prediction, and its ability to capture dynamic changes in customer behavior is limited. Customer churn prediction is a typical scenario with imbalanced positive and negative samples and involves user privacy protection. Further research is needed on the handling of the imbalanced sample problem. Future work will focus on the following areas: introducing time-series analysis techniques to build dynamic predictive models capable of capturing changes in customer behavior over time; exploring the application potential of deep learning methods, particularly recurrent neural networks and long short-term memory networks, in customer churn prediction; combining cutting-edge technologies such as diffusion models to generate synthetic data to address sample imbalance while protecting user privacy; and developing real-time prediction systems to enable enterprises to promptly identify and intervene in potential customer churn behaviors. These research directions will further enhance the practicality and commercial value of customer churn prediction models, providing stronger technical support for enterprise customer relationship management.

References

[1] Hu Tianjie. Feature extraction method that preserves nearest neighbor subspace[J]. Computer Applications and Software, 2024.

[2] Zheng Tong, Shen Ya, Zhang Lijie. Evaluation of digital maturity of Chinese garment enterprises[J]. Wool Textile Technology, 2024.

[3] Chou Fan. Comparative analysis of three velocity field models in CORS reference station time series[J]. Surveying and Mapping Technology and Equipment, 2024.

[4] Yang Guoan, Wang Zhigang. Clustering method using multi-label information[J]. Journal of Tianjin University of Technology, 2024.

[5] Yang Ze. Application of K-means clustering algorithm with PSO integration in financial analysis of colleges and universities[J]. Information and Computer (Theoretical Edition), 2024.

[6] Cao Guilin, Yang Xuliang, Wang Ruofan. Bank customer churn analysis based on machine learning[J]. Journal of Shandong Business Vocational College, 2024.

[7] Huang Junping. Network Intrusion Detection Based on Improved K-means Data Clustering Algorithm [J]. Journal of Chengdu Institute of Technology, 2024.

[8] Jiang Xiaowei, Zhang Wenjin, Liu Ling. Research Progress of Machine Learning Methods in Health Monitoring of Composite Material Structures [J]. Aviation Manufacturing Technology, 2024.

[9] Li Hengbo, Liu Jingchao, Wu Ketong. Image Segmentation Based on Improved K-means Algorithm [J]. Modern Computer, 2024.

[10] Gao Haibin. A K-means Clustering Algorithm Integrating Crow Search Algorithm [J]. Journal of Xinxiang University, 2024.

[11] Sun Lin, Liu Menghan. K-means Clustering Based on Adaptive Cuckoo Optimization Feature Selection [J]. Computer Applications, 2024.

[12] Zhang Fan, Gao Shan. Abnormal Packet Detection Method of Modbus TCP Protocol in Industrial Control Network [J]. Microcomputer Applications, 2024.