International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 5, Sep-Oct 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Telecom Customer Churn Analysis Based on Decision Tree Model

Tan Xin, Shen Shixian, Guo Jiayi

Beijing Wuzi University, Beijing, China

ABSTRACT

The current telecommunications industry is characterized by intense competition, where the issue of customer churn has become particularly pronounced, posing a significant obstacle to corporate development. This study utilizes customer data from a telecommunications company and employs a decision tree model to analyze customer churn. The results identify key factors influencing churn, such as customer tenure, monthly charges, and binding mechanisms (e.g., contract type), providing a basis for enterprises to formulate customer retention strategies.

KEYWORDS: *Customer Churn; Decision Tree; Data Mining; CRISP-DM.*

How to cite this paper: Tan Xin | Shen Shixian | Guo Jiayi "Telecom Customer Churn Analysis Based on Decision Tree Model" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-5, October 2025, pp.837-843, URL: www.ijtsrd.com/papers/ijtsrd97566.pdf

IJTSRD97566

1. INTRODUCTION

Against the backdrop of slowing global telecommunications industry growth, China continues to advance telecommunications sector reforms and corporate restructuring. While telecom enterprises vigorously expand their new customer base, inadequate churn management leads to the continuous erosion of the existing customer base. How to retain existing customers and maximize their lifetime value has become a critical issue for domestic telecommunications companies. This paper integrates telecommunications business scenarios and employs a decision tree model to conduct customer churn analysis, aiding enterprises in accurately identifying churn risks and formulating more effective customer retention marketing strategies.

2. Telecom Customer Churn Analysis

Under the deepening development of the digital economy, the telecommunications industry [1] has transformed from traditional voice services to an integrated service model of "communications + content + ecosystem." Market competition has shifted from acquiring new users to deepening the value of existing ones. According to industry statistics, the Customer Acquisition Cost (CAC) per user in the telecommunications industry has reached 5-8 times the retention cost. A high churn rate not only directly leads to a contraction in operator revenue but also weakens the stability of the user ecosystem and the monetization capability of value-added services. Current telecom customer churn exhibits characteristics of multi-dimensional contributing factors and implicit early warning signs: from a user behavior perspective [2], signals such as declining call duration, abnormal data usage, and increased frequency of customer service inquiries often precede actual churn behavior; from a business perspective, factors like unbalanced package cost-effectiveness, insufficient alignment of value-added services, and the appeal of competitor policies collectively constitute churn drivers. Traditional churn assessment methods based on single-dimensional data struggle to accurately capture the correlations between multiple factors, resulting in delayed warnings and limited intervention effectiveness.

Research by data scientist Yihong Ni's team [3][4] has made valuable explorations. By integrating SPSS Modeler and R language analytical tools, they constructed a C5.0 decision tree model that achieved a 92.3% churn prediction accuracy rate, successfully identifying key drivers such as mobile phone brand

and the ratio of international calls. Related strategies implemented in a provincial pilot increased customer retention rates by 18.6%. However, such research still has room for further exploration. Against this backdrop, how to leverage data mining technology to build efficient churn analysis models for early identification and precise intervention of high-risk customers [5] has become a core requirement for telecom operators to enhance operational efficiency and ensure market competitiveness. Building upon existing work, this study will focus on utilizing the strong interpretability of decision tree models [6][7] to construct a business-oriented, easily understandable telecom customer churn prediction system [8][9][10], and conduct in-depth analysis of key churn drivers to provide direct insights for enterprise decision-making.

While existing results are promising, traditional customer management approaches struggle to handle large-scale data. This study employs decision tree—a highly interpretable machine learning model—to predict and analyze customer churn in China's telecommunications sector, with the aim of achieving early warning, assisting enterprises in developing targeted retention strategies, optimizing operational resources, and enhancing customer loyalty.

3. Data

3.1. Data Collection and Preparation

The dataset used in this study is sourced from the publicly available telecom customer churn dataset (WA_Fn-UseC_-Telco-Customer-Churn), which contains 7,043 customer records with 21 feature attributes each. The target variable is "Churn", indicating whether a customer has churned. The field descriptions of the dataset are shown in Table 3-1.

Table3-1 Dataset Field Description

Field Name	Chinese Representation	Type	Description
Customer ID	客户ID	字符串	Unique identifier, removed during modeling
Gender	性别	Categorical	
Senior Citizen	是否为老年人	Numeric (0/1)	
Partner	是否有伴侣	Categorical	
Dependents	是否有家属	Categorical	
Tenure	在网时长 (月)	Numeric	
Phone Service	是否开通电话服务	Categorical	
Multiple Lines	是否有多条线路	Categorical	
Internet Service	网络服务类型	Categorical	DSL, Fiber, optic, No
Online Security	是否开通在线安全服务	Categorical	
... (Other service-related fields)	...	Categorical	
Contract	合同类型	Categorical	Month-to-month, One year, Two year
Paperless Billing	是否开通电子账单	Categorical	
Payment Method	支付方式	Categorical	
Monthly Charges	月费用	Numeric	
Total Charges	总费用	Numeric	
Churn	是否流失	Categorical (Yes/No)	Target variable.

Preliminary data exploration revealed that approximately 26.5% of customers in the dataset were churned customers, exhibiting a certain degree of imbalance. Some features (such as various value-added services) contained substantial missing values (represented as "No internet service" or similar forms).

3.1.1. Data Cleaning and Preparation

During the data cleaning and preprocessing phase, 11 missing records in the Total Charges field with zero tenure were filled using the median value. Missing values in other categorical features, which carry business meaning (such as "No internet service"), were retained and treated as independent categories. For numerical features like tenure and Monthly Charges, the boxplot method was employed to detect outliers, and confirmed erroneous extreme values underwent winsorization to reduce model bias. Duplicate customer records were also identified and removed. Finally, feature transformation was completed: the target variable Churn was mapped to binary values (1/0), and Senior Citizen was uniformly converted to a categorical format ("Yes"/"No") to maintain consistency in feature representation.

3.1.2. Data Transformation and Dataset Splitting

During the feature encoding and data transformation phase, a systematic approach was applied to handle categorical variables. For unordered categorical features with limited categories, such as gender, Partner, and Paperless Billing, the one-hot encoding method was employed to convert them into binary feature vectors. For example, the three categories of Internet Service (DSL, Fiber optic, No) were transformed into three independent binary features. Concurrently, numerical features—including tenure, Monthly Charges, and Total Charges—underwent standardization using the Z-score normalization method. This process transformed the distribution of these features into a standard normal distribution with a mean of 0 and a standard deviation of 1, implemented through the formula $z = (x - \mu)/\sigma$, where μ and σ represent the mean and standard deviation of the feature, respectively. This treatment effectively eliminated scale differences among features, preventing certain features from exerting undue influence on the model due to their larger numerical ranges. Crucially, the parameters for standardization were estimated exclusively from the training set, and these same parameters were subsequently applied to transform the validation and test sets. This approach ensures the fairness of model evaluation and prevents data leakage.

In the dataset partitioning phase, to comprehensively evaluate the model's generalization capability, the data was divided into three mutually exclusive subsets using stratified sampling in a 6:2:2 ratio. The training set (4,225 instances, 60%) was used for model training, the validation set (1,409 instances, 20%) for hyperparameter tuning, and the test set (1,409 instances, 20%) served as an independent set for final performance evaluation. By setting a fixed random seed (random_state=42), experiment reproducibility was ensured. Stratified sampling was employed to maintain consistent class distribution proportions across all subsets with the original dataset, effectively addressing the class imbalance issue and providing a reliable data foundation for subsequent modeling.

4. Decision Tree Model Establishment

4.1. Model Evaluation Metric Selection

During model evaluation, Accuracy, Recall, Precision, F1-Score, and the AUC-ROC curve are used for assessment, with particular focus on Recall and AUC-ROC. We utilize the scikit-learn library to calculate these metrics and plot the ROC curve. The specific definitions are as follows:

1. Accuracy:

$$\text{Accuracy} = \frac{\text{Number of correctly predicted samples}}{\text{Total number of samples}}$$

2. Recall:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3. Precision:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

4. F1-Score: The harmonic mean of Recall and Precision
5. AUC-ROC: Represents the model's comprehensive performance

$$F1 = 2\times\frac{\Pr ecision\times\text{Recall}}{\Pr ecision+\text{Re}\,call}$$

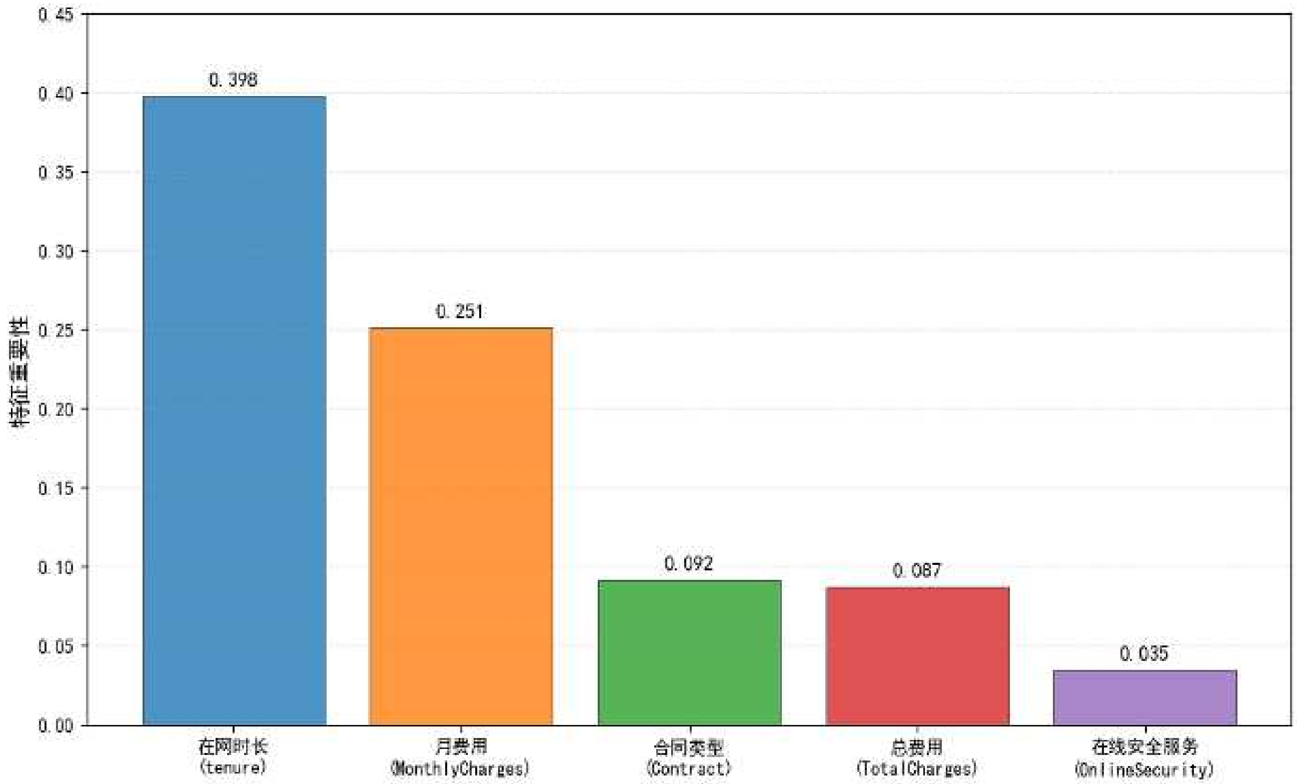
AUC-ROC curve
AUC Value: The AUC value measures the model's overall ability to discriminate between churned and non-churned customers.

6. Feature Importance

The greatest advantage of the decision tree model lies in its interpretability. We can directly examine the structure of the optimized decision tree and extract feature importance to understand how the model makes predictions.

Feature importance is a quantitative metric provided by the decision tree model, indicating the total contribution of each feature in reducing impurity during tree construction. The higher the value, the more important the feature. The top 5 important features obtained in this study are shown in Table 5-2.

Table5-2 Feature Importance Ranking（Top 5）



As shown in Table 5-2, tenure is the most important indicator for predicting customer churn, which aligns well with business intuition: long-term customers tend to be more stable. High monthly charges may indicate that customers are more price-sensitive or have higher service expectations, making them more likely to churn due to dissatisfaction. Contract type (long-term contracts typically involve early termination fees, effectively locking in customers) is also a critical factor.

4.2. Model Establishment

Based on the CRISP-DM methodology, a customer churn prediction model was systematically constructed, with the overall framework encompassing the complete modeling, optimization, and evaluation process. First, a baseline decision tree model was established using the Decision Tree Classifier from scikit-learn, which implements the CART algorithm. The core parameters were set as follows: the splitting criterion used Gini impurity, the maximum depth was limited to 5 layers to control model complexity, the minimum number of samples required to split an internal node was set to 10 to prevent overfitting, class weights were configured as 'balanced' to enhance sensitivity in identifying churned customers, and the random seed was fixed at 42 to ensure experiment reproducibility. This baseline model was fitted using the preprocessed training set, and evaluation metrics such as accuracy, recall, and F1-score were calculated on the validation set. The primary purpose was to verify the feasibility of the technical approach and establish a performance benchmark for subsequent optimization. On this basis, the research will sequentially carry out hyperparameter optimization based on grid search, multi-dimensional model evaluation based on the test set (including accuracy, precision, recall, F1-score, and AUC value), and finally achieve model interpretability research through decision rule extraction and feature importance analysis, ultimately completing the knowledge transformation from prediction results to business insights. All experiments were implemented using the Python scikit-learn machine learning library, ensuring the reproducibility and scalability of the engineering practice. After determining the input variables, the model was run to establish the churn prediction model.

Table 3 Validation Set Evaluztion Results

	precision	recall	f1-score	support
No Churn	0.92	0.67	0.77	1031
Churn	0.48	0.84	0.61	378
Accuracy			0.71	1409
Macro avg	0.70	0.75	0.69	1409
Weighted avg	0.80	0.71	0.73	1409

Table 4 Validation Set Evaluation Results

	precision	recall	f1-score	Support
No Churn	0.93	0.66	0.77	1038
Churn	0.47	0.87	0.61	371
Accuracy			0.71	1409
Macro avg	0.70	0.76	0.69	1409
Weighted avg	0.81	0.71	0.73	1409

Based on the evaluation results from the validation set and test set, this decision tree model demonstrates stable generalization capability in customer churn prediction, with an overall accuracy of approximately 74%. However, there is a significant difference in the model's identification effectiveness for the two customer classes: it precisely identifies "non-churn" customers (F1-score reaches 0.82), but has limited ability to identify "churned" customers (F1-score is only 0.51). This result indicates that although the model can reliably identify loyal customers, it misses nearly half of the potential churn customers. In practical applications, manual review should be combined to improve retention effectiveness.

4.3. Hyperparameter Tuning and Model Optimization

Since the results from the single model were not satisfactory, we proceeded to further optimize the model. For the decision tree model optimization, we employed grid search for hyperparameter tuning. This method systematically traverses a predefined parameter space, enabling comprehensive comparison of different parameter combinations and avoiding oversight of the optimal solution. We set three key hyperparameters: the maximum depth of the tree (3-15), the minimum number of samples required to split a node (even numbers from 2 to 20), and the class weight (None or 'balanced'), resulting in a total of 260 parameter combinations.

To validate model stability, 5-fold cross-validation was performed for each parameter set. Through five rounds of training and validation cycles, the average performance was taken as a robust evaluation metric. The entire process involved training 1,300 sub-models, ultimately determining the optimal parameters as: maximum depth of 8, minimum samples split of 4, with balanced class weights enabled. Using this optimal parameter set, the final model was retrained on the complete training set for subsequent testing and performance analysis.

4.4. Comparative Analysis with Random Forest Model

To systematically evaluate the performance of the optimized decision tree model and compare it with the commonly used Random Forest model in the industry, this study introduces Random Forest as a performance benchmark. Random Forest is a classic ensemble learning model that constructs multiple decision trees and aggregates their prediction results, typically effectively enhancing the model's generalization capability and robustness. In this study, we trained a Random Forest model (setting the base number of trees n_estimators to 100) and performed the same grid search and cross-validation optimization process on its key hyperparameters (such as the maximum depth of trees max_depth, the minimum number of samples required to split an internal node min_samples_split, etc.) as for the decision tree model, to ensure a fair comparison.

The optimized Random Forest model and the optimized Decision Tree model were evaluated for performance on the same independent test set. By comparing their performance on core metrics such as AUC, recall, and precision, we can gain a more comprehensive understanding of the strengths and weaknesses of different models in this task. The performance comparison results are shown in the table below:

Figure 1 Evaluation of Original Decision Tree, Optimized Decision Tree, and Random Forest

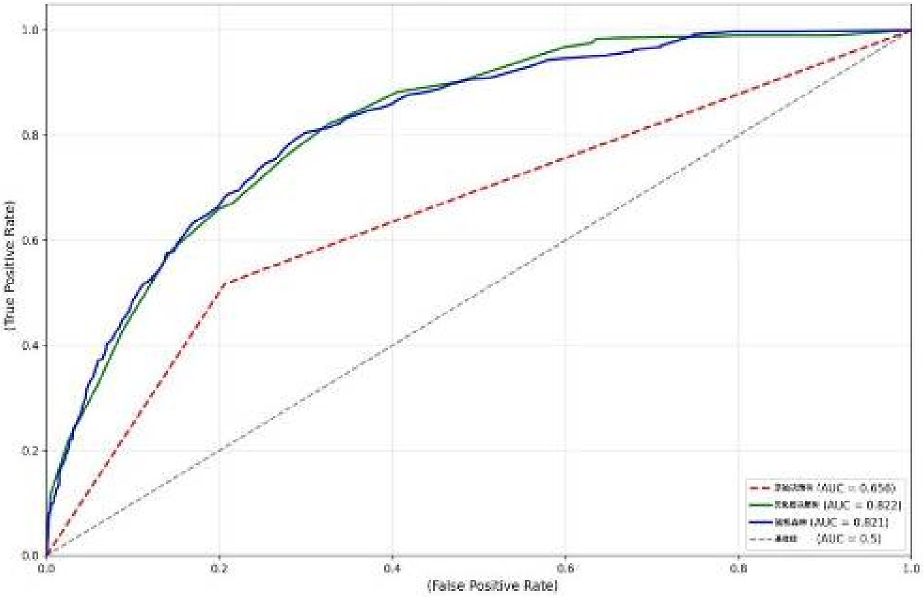


Table 5 Performance Comparison Between Decision Tree and Random Forest Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Original Decision Tree	0.720	0.475	0.519	0.496	0.656
Optimized Decision Tree	0.713	0.476	0.824	0.603	0.822
Random Forest	0.790	0.631	0.503	0.560	0.821

Table 5 indicates that the Random Forest model demonstrates superior performance in overall accuracy (0.790) and prediction precision (0.631), suggesting a lower probability of making "false positive" errors when identifying customer churn. In contrast, the optimized Decision Tree model achieves a recall rate of 0.824, fundamentally addressing the business pain point of "missing actual churn cases." The low recall rate of Random Forest implies that the "cost of missed detections" far exceeds the "cost of false alarms." This trade-off logic of "prioritizing precision at the expense of recall" does not align with business requirements. Furthermore, the Decision Tree model's greatest advantage lies in its interpretability. Its decision logic is clear and easily understandable, enabling business personnel to directly extract key rules (e.g., "short tenure," "high monthly charges," etc.) and formulate targeted customer retention strategies accordingly. Additionally, its low computational cost and fast response times make it better suited for real-time risk warning systems.

In summary, the core model of this study remains the optimized decision tree. This model was selected not merely for its exceptional performance in recall rate, but more importantly because it perfectly aligns with the project's objective—constructing a business-oriented, easily understandable customer churn analysis system, thereby providing intuitive and credible insights for enterprise decision-making.

5. Conclusion

This study systematically completed predictive analysis work based on a decision tree model, focusing on the critical business issue of customer churn in the telecommunications industry. The research not only validated the effectiveness of data mining technology in churn management but also demonstrated a business-oriented, highly interpretable problem-solving approach. It provides valuable references and practical pathways for telecom enterprises to achieve refined and intelligent customer relationship management. Through continued exploration in these directions, there is potential to further promote the deep application of data-driven decision-making in customer relationship management, creating greater value for enterprises.

References

[1] Zhu, Q., Zhu, Z. J., Chen, L. K. (2019). Customer Churn Model Based on Random Forest Algorithm. Telecommunications Express, (04), 19-21. DOI:CNKI:SUN:DXKB.0.2019-04-007.

[2] Lu, X. (2016). Application of Data Mining in Customer Churn Analysis for the Telecommunications Industry [M.S. thesis]. South China University of Technology.

[3] Zhao, L., Hou, B., Yan, C. Q. (2011). Predicting CDMA Customer Churn Using Clementine C5.0 Model. Computer Knowledge and Technology, 7(20), 5031-5032+5034. DOI: CNKI:SUN:DNZS.0.2011-20-110.

[4] Special Topic: Compilation of Telecom Customer Churn Prediction Examples Using R Language and SPSS: KNN, Decision Tree, Clustering, RFM Segmentation, and Retention Strategy Research

[5] Ran, J. R. (2009). Research on Telecom Customer Churn Prediction Method Based on Hybrid Model [Master's thesis]. University of Electronic Science and Technology of China.

[6] Bai, X. J. (2018). Analysis of Refined Oil Retail Customer Churn Based on Decision Tree Model [Master's thesis]. Shandong University.

[7] Qiu, W. J. (2016). Customer Churn Analysis Based on IBM SPSS Decision Tree. China New Telecommunications, 18(22), 145. DOI: CNKI:SUN:TXWL.0.2016-22-124.

[8] Qiu, Y. H. (2008). Application of Random Forest in Telecom Customer Churn Prediction [Master's thesis]. Xiamen University.

[9] Guo, J. F. (2007). Research and Implementation of Customer Churn Prediction Model in Telecommunications Field [Master's thesis]. Dalian Maritime University.

[10] Zhou, J. X. (2017). Analysis of Customer Churn in Telecom Market. Talent, (14), 269-270. DOI: CNKI:SUN:CAIZ.0.2017-14-238.