Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 22, No. 2, 2026

Data-Driven Mobile Network User Satisfaction Analysis: An Optimization Method Based on Multi-Model Fusion

Xiangcheng Xiao *

School of Mathematics and Physics, University of South China, Hengyang, 421101, China

* **Corresponding author**: (Email: x2833363443@163.com)

Abstract: With the rapid development of mobile communication technology, improving user satisfaction has become a core focus for mobile network operators. In an era of ubiquitous connectivity and homogeneous services, understanding and enhancing user experience is crucial for maintaining competitive advantages. Traditional methods to improve user satisfaction, such as responding to complaints and resolving specific issues, have become insufficient with the expansion of user bases and the diversification of mobile services. This study proposes a data-driven approach to analyze the factors influencing mobile user satisfaction in Beijing, focusing on voice and internet services. We combine feature engineering, decision tree models, and ensemble learning techniques to predict and quantify user satisfaction. Specifically, we use the CART (Classification and Regression Tree) model to extract feature importance, and integrate Random Forest and XGBoost models to further improve prediction accuracy through hyperparameter optimization and model fusion. This method combines structured data with unstructured text data, including user descriptions and service notes, to deeply explore the core factors of user experience. Experimental results show that multi-model fusion significantly improves prediction accuracy, with factors such as GPRS traffic, monthly usage, and network issues identified as the main drivers affecting user satisfaction. This study provides valuable in sights for mobile network operators to optimize services and enhance customer experience. Additionally, the proposed method can be extended to other regions and applied to multiple industries where customer satisfaction is critical.

Keywords: User Satisfaction; Random Forest; Feature Engineering; Classification and Regression Tree.

1. Introduction

1.1. Background Introduction

With the rapid development of mobile communication technology, improving user satisfaction has become a key focus for mobile network operators. In this era of ubiquitous connectivity and homogeneous services, understanding and enhancing user experience is crucial for maintaining competitive advantages. Traditional methods to improve user satisfaction, such as responding to complaints and resolving specific issues, have become insufficient with the growth of user bases and the increase in types of mobile services. Data-driven approaches have become increasingly important in customer satisfaction analysis, providing operators with a powerful tool to understand customer needs and improve services [1-2].

Previous studies have emphasized the importance of customer experience management, and big data analytics has been widely used to predict and improve customer satisfaction across various industries. Methods such as decision trees and ensemble learning algorithms have been extensively applied in customer satisfaction prediction, and these models have shown good performance. For example, some studies have used decision tree algorithms to predict the satisfaction of telecom customers, while others have explored the use of hybrid machine learning models.

In mobile communication networks, factors such as network coverage, call quality, and internet speed significantly affect user satisfaction. Recent studies have utilized feature selection techniques, such as PCA and ensemble learning, to identify key factors influencing satisfaction. Furthermore, research has shown that integrating customer feedback data into prediction models can improve the accuracy of user satisfaction prediction [3-4]. With the continuous increase of unstructured data (such as customer descriptions and feedback), the demand for effective text mining and analysis techniques has become more urgent.

As mobile operators continue to improve their services, integrating multiple prediction models (such as CART decision trees, Random Forest, and XGBoost) has been proven to enhance prediction accuracy and robustness. However, how to optimize these models and effectively integrate them to achieve the best results remains a challenge. Hyperparameter optimization methods, such as grid search and cross-validation, have been used to adjust model performance.

1.2. Research Objectives and Significance

This study aims to explore a data-driven approach that combines feature engineering, decision trees, and ensemble learning techniques to predict and quantify mobile user satisfaction. Focusing on voice and internet services, these techniques are applied to analyze the key factors influencing the satisfaction of mobile users in Beijing [5].

The contribution of this study lies in combining structured data with unstructured text data, using TF analysis and word cloud visualization to quantify subjective factors affecting satisfaction. By integrating multiple models, this method provides a more comprehensive and accurate prediction of customer satisfaction in the mobile communication industry.

1.3. Innovations

Text Data Processing: Quantitatively analyze text data such as user descriptions through TF methods, and visualize implicit factors affecting user satisfaction with word clouds, adding a new dimension to customer experience analysis [6].

Ensemble Learning Model Optimization: Combine the

advantages of CART decision trees, Random Forest, and XGBoost, and significantly improve the stability and prediction accuracy of the model through soft voting weighted fusion.

Innovation in Feature Engineering: Adopt PCA dimensionality reduction and feature merging methods to reduce data dimensions, improve computational efficiency, and enhance the generalization ability of decision tree models through post-pruning optimization.

2. Method

2.1. Model Selection and Optimization

In terms of model selection, this study adopts a combination of three machine learning models—CART decision tree, Random Forest, and XGBoost—with subsequent optimization.

CART Decision Tree

The CART model is used to extract feature importance, and post-pruning (Cost Complexity Pruning, CCP) is applied to improve the model's generalization ability. Post-pruning optimizes the tree structure by minimizing cost complexity to avoid overfitting [7-8].

2.2. Hyperparameter Optimization and Model Fusion Strategy

Key hyperparameters optimized include the number of decision tree splits and learning rate. Finally, the prediction results of the three models are combined through soft voting weighted fusion to improve overall prediction accuracy and model robustness.

2.3. Model Evaluation and Validation

To verify the model's performance, the dataset is divided into a training set (70%) and a test set (30%). The training set is used for model training, while the test set is used to evaluate the model's prediction ability. Metrics such as precision, recall, and F1-score are calculated to assess the performance of different models on the test set.

Subsequent experimental results show that prediction accuracy is significantly improved after fusing models through ensemble learning. Especially when dealing with complex features and high-dimensional data, the ensemble model exhibits good generalization ability. Experimental results indicate that the model's test set precision ranges between 0.4 and 0.6, with a recall of 0.65 and an F1-score of 0.60. This shows that the model can effectively identify and classify most positive cases while achieving a good balance between precision and recall.

3. Experiment

3.1. PCA Dimensionality Reduction

A prediction model based on decision trees is established, and an ensemble learning model is introduced. To improve the model's prediction performance, the processed data is further optimized. For the dimensionality-reduced data, feature reconstruction and feature selection are performed, and the decision tree model is pruned to enhance its generalization ability. Subsequently, Random Forest and XGBoost classification models are built to fit the selected dataset. Meanwhile, hyperparameters of the models are optimized through grid search and five-fold cross-validation. Finally, the three models are fused using the soft voting weighted method to further improve the accuracy of model prediction.

The target values of the model include 8 scoring items: "voice call clarity", "overall voice call satisfaction", "overall mobile internet satisfaction", "network coverage and signal strength", "mobile internet stability", "voice call stability", "mobile internet speed", and "network coverage and signal strength". Before selecting features, the dataset needs further processing. This study adopts two methods—PCA dimensionality reduction and feature merging—for feature reconstruction.

Principal Component Analysis (PCA) is an unsupervised multivariate statistical analysis method. Its main idea is to map n-dimensional features to k-dimensional features ($k < n$), where these k-dimensional features are called principal components, reconstructed based on the original n-dimensional features. In this study, all original n-dimensional features of the dimensionality-reduced data are reduced to one dimension.

3.2. Feature Merging

Similar features are summed and merged into a single category based on descriptive fields. For example, features in Dataset 1 (such as no mobile signal, signal but unable to make calls, sudden call disconnection, noise during calls, unclear audio, intermittent calls, cross-talk, and one party unable to hear during calls) are summed to form a new feature "total number of network issues".

3.3. Dataset Splitting

Using the model_selection. train_test_split method from the scikit-learn library, the target values and feature values in the original dataset are divided into two parts at a ratio of 7:3 (training set and test set) by setting a random seed and ratio, which are used for subsequent model training.

1) Decision Tree Post-Pruning (Cost Complexity Pruning, CCP)

To further improve the generalization ability of the decision tree and reduce the risk of overfitting caused by excessive tree growth, post-pruning is performed on the decision tree. Post-pruning involves first generating a complete decision tree from the training set and then pruning and simplifying the tree. CCP optimizes the decision tree by minimizing the following loss function:

$$R_\alpha(T) = R(T) + \alpha|T| \quad (1)$$

Where $|T|$ is the number of leaf nodes in the tree, N is the total number of samples, N_i is the number of samples in the i-th leaf node, and $R(T)$ is the loss function of the i-th leaf node.

α is a pending coefficient: the larger α is, the greater the impact of the number of leaf nodes on the loss function, and the more likely the pruned decision tree is to choose a tree with lower complexity (i.e., fewer leaf nodes). The smaller α is, the smaller the impact of the number of leaf nodes on the loss function, and the more likely the pruned decision tree is to choose a tree with higher complexity (i.e., more leaf nodes). Therefore, α controls the influence of prediction error and tree complexity on pruning.

To determine the optimal value of α, a learning curve is first plotted to select the optimal range. Using Python programming, test samples and training samples are substituted into the model to plot the accuracy of the decision tree on training and test samples under different α values. Taking two scoring items— "overall voice call satisfaction"

in Dataset 1 and "overall mobile internet satisfaction" in Dataset 2—as examples:

The optimal accuracy range for both is [0.0000 ~ 0.0125]. To improve model accuracy, α values within this optimal range are selected in subsequent hyperparameter optimization. This study completes the CCP pruning operation on the CART decision tree and plots the post-pruning visualization graph. Due to space limitations, only part of the post-pruning visualization for overall voice call satisfaction is shown here. The accuracy of the test set after pruning is effectively improved, and overfitting is significantly alleviated.

3.4. Model Selection and Optimization

First, Random Forest is used to generate a set of base learners, which are then combined using a specific strategy. Generally, decision trees are used as base learners, which are inherently weak learners but become strong learners with high prediction performance after integration. In essence, Random Forest is the majority voting result of multiple decision trees combined together.

Algorithm Steps:

Train the model using training data: Extract the training set from the data sample set through bootstrap sampling for B rounds, resulting in B mutually independent training sets.

Feature selection: Select features based on the important feature selection described above.

Train one model using each training set, resulting in a total of B models.

Obtain the classification result through voting and output it as the final prediction.

Then, the XGBoost model ensemble learning algorithm is adopted to upgrade weak learners to strong learners. Based on the idea of Gradient Boosting Decision Trees (GBDT), XGBoost optimizes the GBDT model. XGBoost is an additive model that serially trains a set of CART decision trees and weighted sums the prediction results to form a strong learner. The commonly used objective function of the model is:

$$\mathrm{Obj}(\theta) = \mathrm{L}(\theta) + \Omega(\theta) \tag{2}$$

Where $\mathrm{Obj}(\theta)$ is the objective function, $\mathrm{L}(\theta)$ is the model training loss used to evaluate model performance, $\Omega(\theta)$ is the regularization term, and $\mathcal{D}$ is the sample space.

XGBoost performs linear weighted integration based on a single tree model, and training is conducted in an additive manner. Its prediction principle is: the prediction value at a certain moment is the sum of the prediction value at the previous moment and the function value at the current moment, with the goal of selecting the optimal function value. Let the training dataset be $\mathcal{D} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i=1}}^{\mathrm{n}}$, and the expression of its objective function is:

$$\mathrm{Obj}^{(\mathrm{t})} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}) + \sum_{\mathrm{k=1}}^{\mathrm{t}} \Omega(\mathrm{f_k}) \tag{3}$$

Where $\hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}$ represents the prediction value of the i-th sample at the t-th iteration, and L is the loss function (softmax loss function is used in this study). The loss function $\mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}) = \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})} + \mathrm{f_t}(\mathrm{x_i}))$, where $\Omega(\mathrm{f_t})$ is the regularization term representing the complexity of the model. In the gradient boosting tree model, $\hat{\mathrm{y}}_\mathrm{i}^{(0)} = 0$, and the objective function is split into:

$$\mathrm{Obj}^{(\mathrm{t})} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})} + \mathrm{f_t}(\mathrm{x_i})) + \Omega(\mathrm{f_t}) \tag{4}$$

Where $\mathrm{f_t}$ represents the function at the t-th iteration. The principle is to minimize the objective function by finding the optimal $\mathrm{f_t}$. The Taylor expansion is used to expand $\mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})} + \mathrm{f_t}(\mathrm{x_i}))$, and the Taylor formula is:

$$\mathrm{f}(\mathrm{x} + \Delta\mathrm{x}) \approx \mathrm{f}(\mathrm{x}) + \mathrm{f}'(\mathrm{x})\Delta\mathrm{x} + \frac{1}{2}\mathrm{f}''(\mathrm{x})\Delta\mathrm{x}^2 \tag{5}$$

Using the Taylor formula, the objective function can be expressed as:

$$\mathrm{Obj}^{(\mathrm{t})} \approx \sum_{\mathrm{i=1}}^{\mathrm{n}} \left[\mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})}) + \mathrm{g_i}\mathrm{f_t}(\mathrm{x_i}) + \frac{1}{2}\mathrm{h_i}\mathrm{f_t}^2(\mathrm{x_i})\right] + \Omega(\mathrm{f_t}) \tag{6}$$

Where $\mathrm{g_i} = \frac{\partial \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})})}{\partial \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})}}$ denotes the first derivative of the loss function for the i-th sample, and $\mathrm{h_i} = \frac{\partial^2 \mathrm{L}(\mathrm{y_i}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})})}{\partial (\hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t-1})})^2}$ denotes the second derivative of the loss function for the i-th sample.

Thus, the objective function can be rewritten as:

$$\mathrm{Obj}^{(\mathrm{t})} \approx \sum_{\mathrm{i=1}}^{\mathrm{n}} \left[\mathrm{g_i}\mathrm{f_t}(\mathrm{x_i}) + \frac{1}{2}\mathrm{h_i}\mathrm{f_t}^2(\mathrm{x_i})\right] + \Omega(\mathrm{f_t}) \tag{7}$$

Where the complexity of the tree model is defined as $\Omega(\mathrm{f_t}) = \gamma\mathrm{T} + \frac{1}{2}\lambda\sum_{\mathrm{j=1}}^{\mathrm{T}} \mathrm{w_j}^2$, and the parameters of this complexity can be used to update the objective function.

When constructing the Random Forest and XGBoost models, initial models are first built with default parameter values to obtain initial evaluation metrics. Then, important hyperparameters of XGBoost (such as max_depth, n_estimators, learning_rate) and Random Forest (such as max_depth, n_estimators, min_samples_leaf, min_samples_split) are adjusted. Among these, learning_rate is the learning rate used to control the iteration speed; n_estimators is the number of iterations (i.e., the number of weak classifiers); max_depth is the maximum depth of the tree used to avoid overfitting; min_samples_split and min_samples_leaf are the minimum number of samples required to split an internal node and the minimum number of samples required for a leaf node, respectively. These parameters can limit the complexity of the model, produce a smoothing effect, and mitigate overfitting.

Grid search and cross-validation for different parameter combinations are time-consuming. To improve parameter optimization efficiency, learning curves are usually plotted to determine the optimal accuracy range for individual parameters. Subsequently, the optimal parameter combination is obtained through grid search combined with five-fold cross-validation within the optimal range of each parameter.

Taking two scoring datasets—overall voice call satisfaction and overall mobile internet satisfaction—as examples:

First, it is necessary to control the maximum depth (max_depth) of a single base learner tree. The fitting ability of a classification tree is proportional to its complexity. Generally, the deeper the tree, the more branches it has, making it more specific but less generalizable. Therefore, limiting the maximum depth can effectively avoid overfitting. Figure 1 shows the changes in errors of the training and test sets with depth for Random Forest and XGBoost.

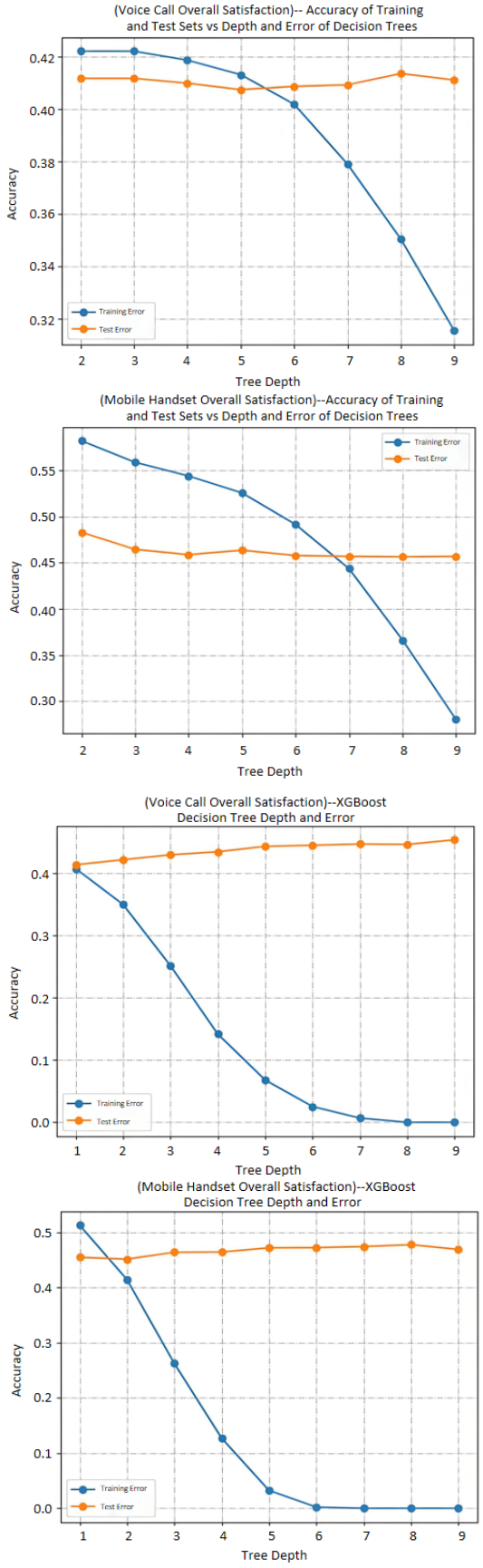

Figure 1. Changes in errors of training and test sets with depth for Random Forest and XGBoost

It can be seen that both models experience severe overfitting as the tree depth increases. Analysis of prediction errors for overall mobile internet satisfaction and overall voice call satisfaction shows that for the Random Forest model, when the maximum depth of a single tree is in the range of 5~7, the test error and training error gradually reach a balance; beyond this depth, the degree of overfitting increases significantly. For the XGBoost model, the balanced depth range of test error and training error for the two scoring datasets is 1~2. Based on the above analysis, the optimal ranges for the maximum depth of a single tree in the prediction models for Random Forest and XGBoost are [5, 7] and [1, 2], respectively. Next, the minimum number of samples required to split internal nodes and the minimum number of samples required for leaf nodes in the Random Forest model are controlled to further mitigate overfitting. Finally, the optimal combination of these three parameters for Random Forest is found through grid search and five-fold cross-validation.

Similarly, to better determine the optimal iteration number range of the XGBoost model and improve its robustness, it is necessary to adjust the learning rate (learning_rate) of XGBoost to control the step size of weight updates in each iteration. Refer to the learning rate error change curve for details.

Next, it is necessary to control the number of base learners (iteration times) in the two ensemble models to find the convergence range of each model under low error to achieve an ideal prediction effect. The error change curve is shown in Figure 2:

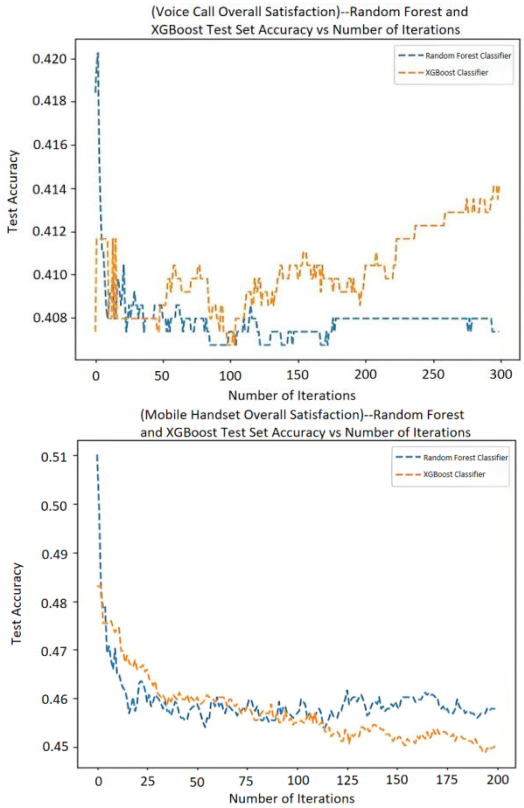

Figure 2. Changes in test errors of Random Forest and XGBoost with the number of iterations

Analysis of the figure shows that for the prediction of voice call and mobile internet overall satisfaction data, Random Forest converges faster to a lower test error than XGBoost. Its minimum test error range appears after 50 iterations, while that of XGBoost is between 100~200 iterations. After determining the optimal iteration number range, the optimal parameter values for the iteration number parameter (n_estimators) of the two models for predicting each scoring dataset are found through grid search and five-fold cross-validation.

Finally, the optimal model is constructed using the selected

optimal parameter combination, and the accuracy of each model before and after parameter tuning is calculated.

3.5. Model Fusion

Based on the three established models, a fusion approach is adopted using soft voting weighted fusion. The specific model results are as follows:

$$P_{final} = \beta_1 P_{CART} + \beta_2 P_{RF} + \beta_3 P_{XGBoost} \quad (8)$$

Where β_1, β_2, β_3 are weight coefficients, P_{CART}, P_{RF}, $P_{XGBoost}$ represent the probabilities of being classified into a certain result in the CART model, Random Forest model, and XGBoost model, respectively. After continuous debugging, the optimal weight combination is obtained as $\beta_1 = 0.2$, $\beta_2 = 0.3$, $\beta_3 = 0.5$. The calculated P_{final} is then normalized to obtain the final classification probability, as shown in Figure 3.

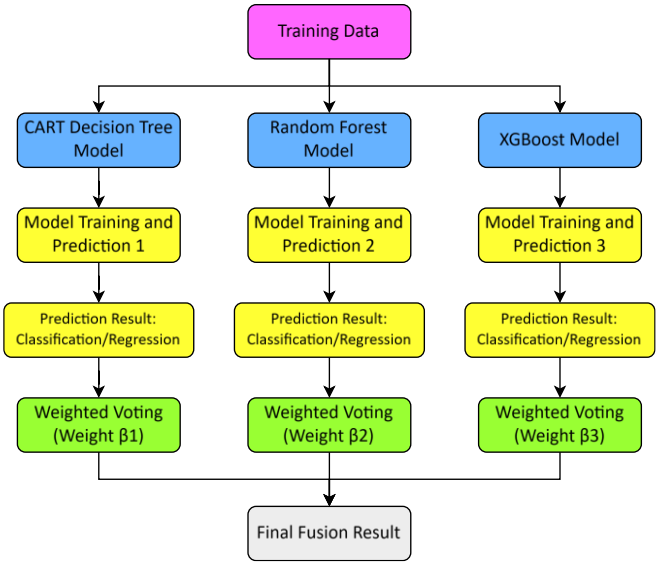

Figure 3. Schematic diagram of model fusion

Subsequently, the test errors of the CART model, Random Forest model, XGBoost model, and the fused model are calculated. The test error after model fusion is minimized, indicating that the prediction results of the fused model are more accurate. The above results show that the final test set accuracy of the model is between 0.4 and 0.6, which is not very high. However, combined with the descriptive statistics of the data in Problem 1, it is found that most indicators in the data have outliers, i.e., customer scoring is highly subjective. Under the same factors, different people have different needs, leading to inconsistent scoring. The recall rate is 0.65 and the F1-score is 0.60, verifying that the model achieves a good balance between precision and recall. It indicates that through continuous learning, debugging, and fusion of the model, the accuracy of the test set is continuously improved. Thus, the final prediction results are in line with reality and reasonable.

4. Conclusion

Research Summary: A data-driven user satisfaction analysis method is proposed, and prediction accuracy is improved through multi-model fusion. A quantitative analysis of the key factors influencing the experience of Beijing Mobile users is conducted, providing a scientific basis for operators to improve service quality.

Future Outlook and Innovations: It is proposed that in future work, more types of data (such as social media comments and real-time network data) will be integrated to further optimize the model and improve prediction accuracy. Explore the combination of artificial intelligence technology and customer service to adjust service strategies in real-time to achieve higher user satisfaction.

Summary of Innovations: Introduction of Text Data Processing: Quantify subjective factors in user satisfaction through TF analysis and word clouds, an innovation that breaks through the traditional structured data analysis framework. Introduction and Optimization of Ensemble Learning Models: Combine the advantages of Random Forest, decision trees, and XGBoost, and improve the stability and prediction accuracy of the model through soft voting weighted fusion. Innovation in Feature Engineering: Effectively handle high-dimensional data through PCA dimensionality reduction and feature merging methods, and improve the generalization ability of the model through post-pruning optimization. Model Promotion Value: The model is not only applicable to Beijing Mobile but also can be promoted to other regions, with practical commercial application value.

References

[1] ZHOU Y, CHENG X. Improving user satisfaction in mobile networks: a review of methods and models [J]. International Journal of Mobile Communications, 2022, 20(4): 527-540.

[2] HUANG S, XIE J, LI R, et al. Research on the impact factors of Beijing mobile user experience based on random forest method [J]. Science and Technology Innovation, 2024(12): 45-50.

[3] QU H, WANG Z, SHENG Z, et al. Signal classification of optical fiber intrusion based on gradient boosting decision tree algorithm [J]. Laser & Optoelectronics Progress, 2022, 59(4): 1-8.

[4] MA T. Design and implementation of student profiling generation system [D]. Shanghai: East China Normal University, 2023.

[5] LAN Z, WANG S, CAO Y, et al. Customer classification for urban gas based on Kmeans++ algorithm and LGBM model [J]. Natural Gas Technology and Economy, 2024, 32(2): 12-18.

[6] WANG H, CHEN D, LI L, et al. Effectiveness evaluation of air combat system based on ensemble learning [J]. Firepower and Command Control, 2023, 48(1): 23-30.

[7] WANG Y, LIU E, HUANG Y. Short-term multi-dimensional load forecasting for integrated energy systems driven by data [J]. Computer Engineering and Design, 2022, 43(8): 2153-2159.

[8] WU M, MA L, YANG A, et al. Application of federated learning in mobile communication network intelligence [J]. Mobile Communication, 2022, 45(3): 15-20.