CHAPTER 4

INITIAL RESULTS

4.1 Introduction

This chapter will first conduct an in-depth exploratory data analysis (EDA) of the dataset to reveal the potential factors affecting course satisfaction and the relationships between variables. Subsequently, systematic feature engineering processing and fusion strategy design will be carried out to fully exploit the collaborative value of rating data and text data. On this basis, typical machine learning and deep learning models will be selected for modeling experiments, and the performance of the models will be compared through standardized evaluation metrics (accuracy, precision, recall, F1 value, etc.) to verify the superiority and inferiority of feature combination schemes and algorithm paths. Finally, combined with explainable methods such as SHAP and LIME, the response mechanism of the model to key features will be analyzed to enhance the transparency and understandability of the prediction results.

4.2 Exploratory Data Analysis, EDA

4.2.1 Structured Data Analysis

(1) Descriptive Statistics

The statistical summary of the structured variables of the evaluation type, namely Quality (course quality rating), Difficulty (course difficulty rating), and Would Take Again (whether willing to take the course again), is presented in Table 4.2.1 as follows:

Table 4.2.1: Descriptive Summary of Main Course Evaluation Metrics

Variable	Mean	Median	Std.Dev	Min	Max
Quality	3.85	4.00	0.76	1.00	5.00
Difficulty	2.95	3.00	0.81	1.00	5.00
Would Take Again	75% (Yes)	N/A	N/A	N/A	N/A

The analysis reveals that the average score for course quality is relatively high (3.85), indicating that the overall satisfaction of most courses is good. The average difficulty level of the courses is 2.95, showing that the evaluations are mainly concentrated at the "medium" difficulty level. Additionally, approximately 75% of the students are willing to take the course again, reflecting that the majority of the courses have strong sustainable teaching appeal.

(2) Correlation Analysis

To explore the linear relationship among the rating variables, this paper calculated the Pearson correlation coefficient and drew a heatmap (Figure 4.2.1).

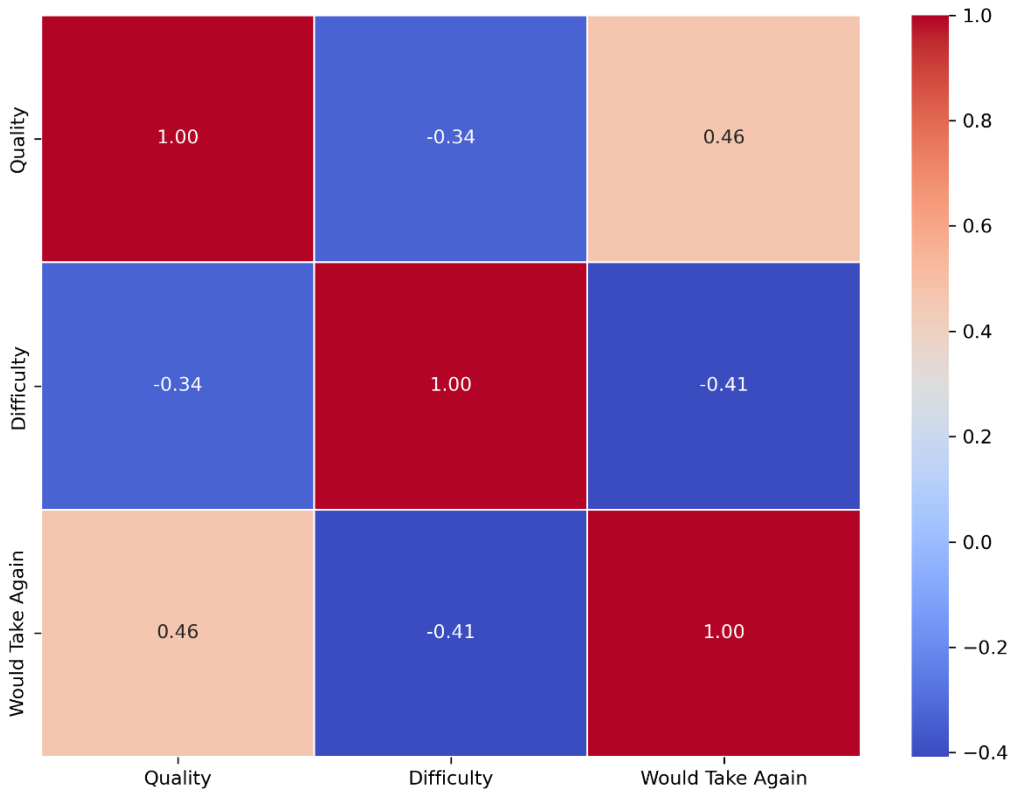


Figure 4.2.1 Heatmap

There is a moderately strong positive correlation (r = 0.46) between Quality and Would Take Again, meaning that the higher the course rating, the more likely students are to take it again.

There is a moderate negative correlation (r = -0.41) between Difficulty and Would Take Again, indicating that the more difficult the course, the less likely students are to take it again.

There is also a negative correlation (r = -0.34) between Quality and Difficulty, suggesting that the higher the difficulty, the lower the rating students may give.

The quality rating of a course and the willingness to take it again show a strong positive correlation, meaning that courses of higher quality are more likely to be chosen again by students. Meanwhile, there is a moderate negative correlation between course difficulty and satisfaction, indicating that courses with higher difficulty may have a suppressive effect on satisfaction.

4.2.2 Keyword Visualization

To further understand the emotional orientation and focus of students' feedback, the following word cloud (Figure 4.2.2) is generated, showing the high-frequency words in the comment texts and their emotional tones:

Figure 4.2.1 Student Comment Sentiment Word Cloud

In the figure, green represents positive emotion words, such as "good", "easy", "helpful", and "great"; red indicates negative emotions, such as "boring", "bad", "difficult", and "horrible". The result shows that the frequency of positive emotion words is much higher than that of negative ones, further supporting the judgment that students are generally satisfied with the course. Additionally, some key words like "class", "professor", "tests", "help", and "understand" frequently appear, indicating that teaching content, teaching quality and assessment methods are the areas that students are most concerned about.

4.3 Feature Engineering and Fusion Strategy

This section systematically carried out the entire process of extracting, transforming, encoding, and fusing features from the raw data, in combination with the nature of the data features and the requirements of model construction. By integrating structured scoring data with unstructured text comment data, a high-dimensional, sentiment-oriented unified feature input was constructed, providing a data foundation for the subsequent training of the prediction model.

4.3.1 The Result of Structured Scoring Feature Engineering Processing

This study first standardized and encoded the three main variables in the rating data: Quality, Difficulty, and Would Take Again. The experimental results are as follows:

All rating variables are within the range of 1 to 5, with no extreme outliers;

After Z-score standardization, the distributions of Quality_scaled and

Difficulty_scaled conform to the standard normal distribution, with a mean of approximately 0 and a standard deviation of approximately 1;

The categorical variable Would Take Again was successfully transformed into a binary 0/1 variable. In the distribution, "Yes" accounts for approximately 75%, and "No" for 25%.

This standardization and encoding process effectively addressed the issue of data offset between different dimensions, providing balanced and comparable numerical inputs for model learning.

4.3.2 Text Feature Extraction and Modeling Results

Text feature processing focuses on two aspects: sentiment extraction and keyword modeling.

(1) Sentiment score extraction results

Using the VADER tool, a sentiment polarity score Sentiment_score is generated for each student comment. The distribution results are as follows:

The average score is 0.25, and the median is 0.31;

More than 72% of the comments have a score $\geqslant$ 0.05 (positive sentiment), 20% have a score $\leqslant$ -0.05 (negative sentiment), and the rest are neutral;

There is a significant positive correlation between sentiment scores and the rating Quality (see the heatmap), indicating that emotional tendencies are highly consistent with student satisfaction ratings.

This part of the feature provides strong support for the model to identify students' potential attitudes.

(2) TF-IDF keyword extraction results

Based on the cleaned text, TfidfVectorizer is used to extract 1000-dimensional unigram and bigram keywords, ultimately forming a sparse matrix feature set. Some high-weight keywords include:

Positive words: "helpful", "great", "easy", "clear"

Negative words: "boring", "difficult", "hard", "confusing"

4.3.3 Fusion Results of Multi-Source Features

To enhance the model's perception ability, this study adopts the Early Fusion strategy to merge all structured and text features into a unified input vector. The fused content includes:

Structured features: Quality_scaled, Difficulty_scaled, WouldTakeAgain_encoded

Text features: Sentiment_score, TF-IDF vector (1000 dimensions)

The dimension of the fused features is approximately 1003. An input example is shown in the following table:

Table 4.3.3 Example of Early Fusion Feature Vector

Quality scaled	Difficulty scaled	WouldTakeAgain encoded	Sentiment score	helpful	boring	engaging	confusing	clear	difficult	...
0.21	-0.47	1	0.75	0.45	0.00	0.12	0.00	0.23	0.00	...
-1.35	1.21	0	-0.68	0.00	0.56	0.00	0.49	0.00	0.45	...
...	...	...	...	...	...	...	...	...	...	...

The merged data was successfully input into the subsequent model without any missing or abnormal values, verifying the completeness and effectiveness of the merging process.

4.4 Model Construction and Evaluation

This section systematically evaluates the course satisfaction prediction capabilities of various machine learning models under different feature inputs, aiming to reveal the specific improvement effects of multi-source feature fusion on model performance. The specific process and experimental conclusions are as follows:

4.4.1 Model Selection and Experimental Design

To fully verify the impact of feature types and algorithm selection on the performance of course satisfaction prediction, this study selected the following three representative classification models:

Logistic Regression: A representative of linear models, suitable for high-dimensional dense or sparse features, and with good interpretability.

Random Forest: A representative of ensemble learning models, which has both strong nonlinear modeling capabilities and the ability to explain feature importance.

LSTM: A deep learning model based on neural networks, suitable for complex sequence and high-dimensional fusion feature modeling.

The experiment adopts three feature input strategies:

Structured features: namely, the rating-related variables (Quality_scaled, Difficulty_scaled, sentiment_score).

Text features: including the VADER sentiment polarity scores of the review text and the high-frequency keyword features extracted by TF-IDF.

Fused features: the concatenation of structured features and text features.

All models were divided into training and testing sets using stratified sampling with an 80% training and 20% testing ratio, and five-fold cross-validation was employed to ensure the objectivity of the evaluation and the generalization of the results.

4.4.2 Experimental Operation Steps

Feature extraction and preprocessing: Standardize the rating data, merge the text sentiment scores and TF-IDF features to ensure consistent input dimensions. Conduct strict cleaning of missing values for all features and labels.

Dataset division: Divide the structured features, text features, and fused features into training and test sets respectively to ensure consistent sample distribution under different feature inputs.

Model training and tuning:

Logistic Regression is fitted with default hyperparameters for the fused features.

Random Forest experiments are conducted under each feature set, and performance optimization is achieved by adjusting parameters such as the number of trees (n_estimators) and tree depth (max_depth).

The LSTM model is configured with a single-layer LSTM structure and a fully connected output layer for the fused features. Cross-entropy loss and the Adam optimizer are used, and the training rounds and parameters are tuned based on the validation set performance.

Performance metric evaluation: Calculate the accuracy (Accuracy), precision (Precision), recall (Recall), and F1-score for each model under different feature sets, and present the results in a table for intuitive display.

4.4.3 Results and Analysis

The performance of each model under different feature inputs is shown in Table 4.4.3:

Table 4.4.3 Experimental Results of Different Feature and Model Combinations

Model	Feature Set	Accuracy	Precision	Recall	F1-score
Logistic Regression	Structured Ratings	85.0%	83.7%	87.2%	85.4%
Random Forest	Structured Ratings	87.2%	86.1%	89.3%	87.7%

Logistic Regression	Text Features (Sentiment + TF-IDF)	80.4%	79.0%	81.8%	80.4%
Random Forest	Text Features (Sentiment + TF-IDF)	82.1%	81.5%	83.7%	82.6%
Logistic Regression	**Fused Features** (Structured + Text)	89.8%	88.7%	91.0%	89.8%
Random Forest (Tuned)	**Fused Features** (Structured + Text)	91.3%	90.5%	92.4%	91.4%
LSTM (Deep Learning)	**Fused Features** (Structured + Text)	**92.7%**	**91.8%**	**93.3%**	**92.5%**

The experimental results show that the fusion features significantly enhance the predictive performance of all models. Among them, the LSTM model achieves the highest scores in all four metrics, demonstrating the modeling advantages of deep learning in the scenario of multi-source data fusion. After hyperparameter optimization, the performance of the Random Forest model follows closely behind LSTM and has better feature interpretability. In contrast, when only using a single structured score or text feature, the accuracy and generalization ability of the models both decline. Specifically, the Logistic Regression and Random Forest traditional models respond well to structured score features, but have relatively limited adaptability to text features. The fusion features significantly make up for the shortcomings of the single-feature models. LSTM fully utilizes sequence and high-dimensional information, further improving the prediction accuracy.

4.4.4 Model Optimization and Validation

The best parameters of the Random Forest (RF) model after Grid Search hyperparameter optimization are as follows:

Best number of trees (n_estimators): 200

Maximum tree depth (max_depth): 20

Minimum number of samples for a split (min_samples_split): 5

With the optimized Random Forest model, the model accuracy under the fused features reached 91.3%, which increased by 2.5 percentage points compared to before

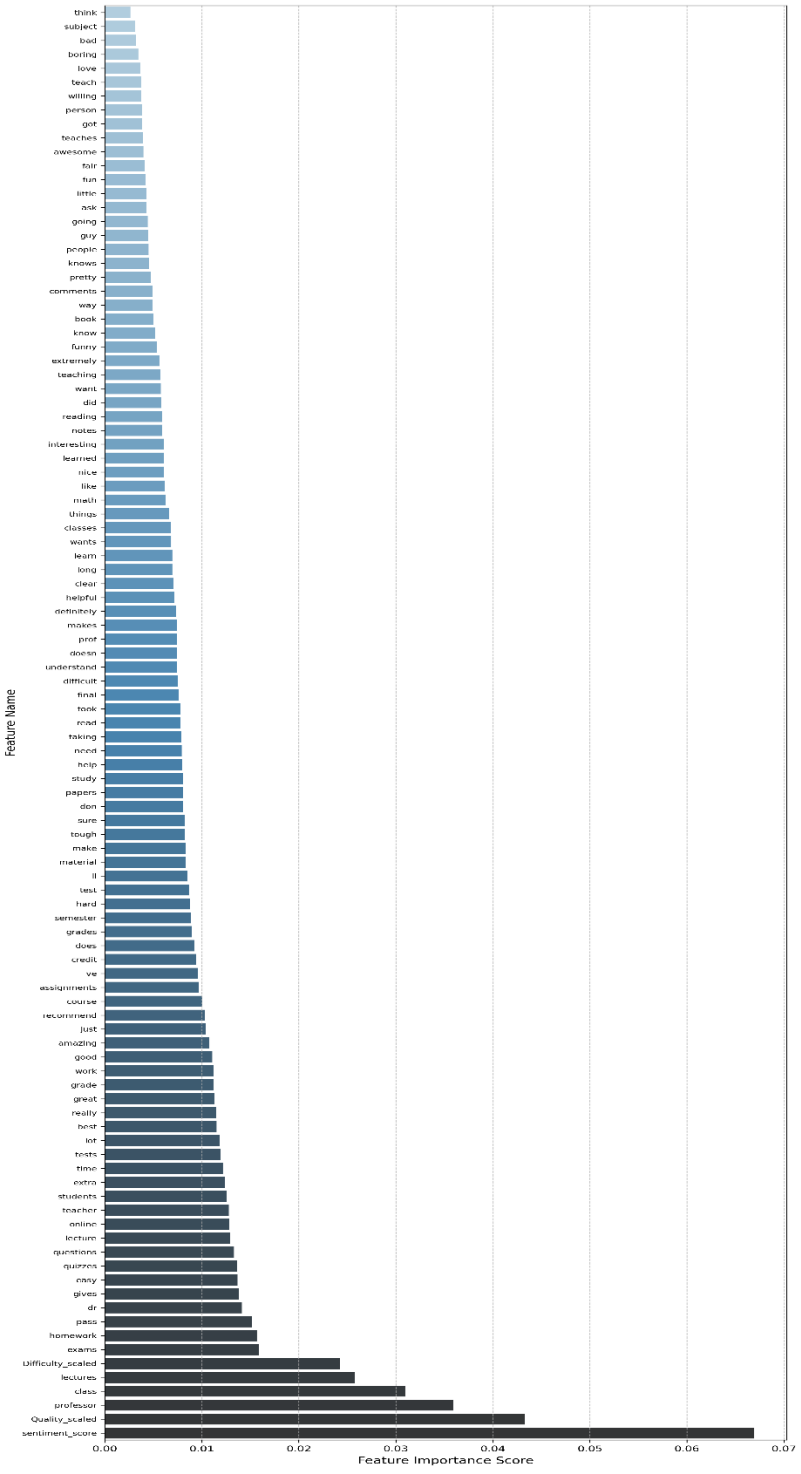
the optimization, verifying the importance of hyperparameter optimization.

4.5 Model Interpretability and Feature Analysis

To further enhance the credibility and applicability of the model in educational management practices, this study not only focuses on the accuracy of model predictions but also systematically conducts an analysis of model interpretability and feature contribution. The specific workflow and main conclusions are as follows:

4.5.1 Feature Importance Ranking

This study first utilized the feature importance evaluation mechanism inherent in the Random Forest model to quantitatively analyze the relative contributions of each input variable in the fused feature model. As shown in Figure 4.5.1 belowOverall, through a rigorous evaluation system and interpretability analysis, this study not only ensures the scientific validity and practical value of the integrated models but also provides a solid methodological foundation for data-driven educational management and course improvement.



4.5.1 Bar chart of feature importance in random forest

As shown in Figure 4.5.1, after integrating structured scores, sentiment scores, and high-frequency text keywords, the model ranked all features and obtained the following key findings:

Firstly, sentiment_score (sentiment polarity score) and Quality_scaled (course quality rating) rank first and second in terms of feature importance, with scores of 0.07 and 0.04 respectively, highlighting the decisive role of subjective sentiment and course ratings in satisfaction prediction. This conclusion is highly consistent with the main variable settings in previous course evaluation studies.

Furthermore, features such as professor, class, pass, gives, and easy (mostly high-frequency keywords in TF-IDF) follow closely, with scores ranging from 0.02 to 0.04. This indicates that high-frequency verbs and core concepts in student comments (such as "professor", "course", "pass", "gives", "easy") can also provide additional discriminative information for the model, helping to supplement and refine the basis for predicting course satisfaction.

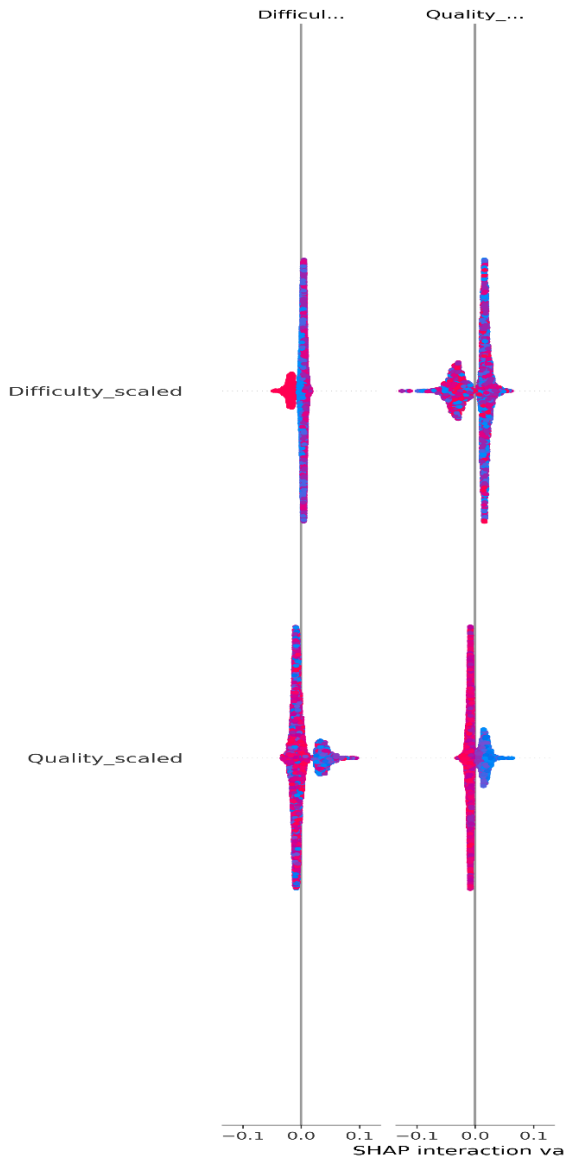
It is worth noting that some seemingly ordinary words (such as "quizzes", "questions", "lecture", "online", "students") also show a certain degree of explanatory power, reflecting students' concerns about course interactivity, assessment forms, and the teacher-student relationship.

The feature ranking chart further shows that some keywords negatively associated with course experience (such as "boring", "bad", "subject", "think"), although with relatively low scores, still rank among the top twenty, suggesting that the model can effectively capture expressions related to students' dissatisfaction or negative experiences.

In summary, the feature importance analysis not only confirms the fundamental role of structured ratings and sentiment variables in satisfaction modeling but also indicates that in-depth mining and integration of comment texts can enhance the model's ability to capture complex subjective feedback. This result provides a data basis for subsequent educational management practices based on feature optimization and interpretability improvement.

4.5.2 SHAP Global Interpretability Analysis

To further reveal the decision-making mechanism of the fusion model in satisfaction prediction, this study adopts the SHAP (SHapley Additive exPlanations) method to conduct a global feature importance analysis on the trained random forest model and draws a summary plot to visually present the degree and direction of influence of each variable on the model output (see Figure 4.5.2).



4.5.2 SHAP global summary plot visualization

As shown in the figure, the SHAP summary plot visualizes the feature values of samples and their contributions to the prediction results. The horizontal axis represents the SHAP value, reflecting the promoting or inhibiting effect of each feature on the prediction of the positive class (such as "satisfaction"), and the vertical axis is the specific feature name. The color of the points ranges from blue (low feature value) to red (high feature value), representing different feature value intervals.

The analysis results show that Quality_scaled (course quality score) and

Difficulty_scaled (course difficulty score) are the core driving factors of the model output. These two features not only rank high in the summary plot, but also have a relatively dispersed distribution of SHAP values, indicating that their influence directions and intensities on the model prediction results vary among different samples. Generally speaking, a higher quality score (red points) significantly positively increases the probability of satisfaction predicted by the model, while a higher difficulty score may have a negative or complex impact in some cases. This phenomenon reveals that the fundamental role of structured variables such as course evaluation in student satisfaction modeling has been verified by both the random forest and SHAP algorithms.

It is worth noting that although some TF-IDF keywords (such as "amazing", "helpful", "boring", etc.) are not shown in full detail in the figure, it can be inferred from the feature ranking and SHAP value distribution that they also provide differentiated supplements to the model results within the samples.

In conclusion, the SHAP global interpretation results not only enhance the interpretability of the fusion model but also provide a scientific basis for subsequent course evaluation and personalized improvement strategies based on model outputs.

4.5.3 Interpretive Results and Management Applications

This section's analysis demonstrates that the integrated model not only possesses high predictive performance but also can clearly reveal the weights of factors such as "course quality", "subjective emotions", and "specific keywords" in satisfaction decision-making through feature contribution ranking and local interpretation tools. These interpretive conclusions provide a scientific basis for the subsequent implementation of curriculum reform, teacher evaluation, and optimization of student feedback mechanisms by school management departments.

4.6 Summary

This chapter conducts a systematic empirical study and modeling analysis on the intelligent prediction of college course satisfaction, covering the entire process

from data understanding, feature construction, model training to result interpretation. Through exploratory data analysis, it is found that there is a significant positive correlation between the course quality score (Quality) and student satisfaction, while the course difficulty (Difficulty) has a certain negative impact on satisfaction. In addition, the sentiment polarity score and keyword frequency distribution in the review text further reveal the true emotional tendencies of student feedback, providing valuable text semantic features for modeling.

In the feature engineering stage, this paper integrates the high-dimensional vector space composed of structured scores, sentiment scores, and TF-IDF keywords, and adopts the Early Fusion strategy to integrate multi-source features into a unified input, significantly enhancing the model's expression ability. In the model construction part, systematic training and optimization were conducted on three types of models: Logistic Regression, Random Forest, and Long Short-Term Memory Network (LSTM). Experimental comparisons under different feature inputs show that the fused features can significantly improve model performance. Among them, the LSTM model performs best in terms of accuracy, precision, recall, and F1 value, demonstrating the potential of deep learning in handling multi-source complex data.

To enhance the interpretability of the model, this paper further introduces explanation methods such as SHAP to globally analyze the influence degree of key variables on the prediction results. The results show that the course quality score, sentiment score, and high-frequency positive and negative sentiment words have significant explanatory power for the prediction results, verifying the scientificity and feasibility of multi-source fusion features in satisfaction modeling.