Proceedings of the 7th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/38/20231913

Improved Random Forest Based on Grid Search for Customer Satisfaction Prediction

Ruixi Luo[1,a,*]

[1]*ssfbc, Xi'an, China*
a. 1141835418@qq.com
**corresponding author*

Abstract: Customer satisfaction is an important influencing factor that affects the survival and development of an enterprise, especially a service-oriented enterprise, then predicting customer satisfaction can achieve a continuous and steady improvement of customer satisfaction of an enterprise. In this paper, we improve the random forest algorithm based on grid search, and use the improved random forest model to predict airline customer satisfaction scores and rank the importance of customer satisfaction influencing factors by minimizing the average root mean square error of the model as the goal of the preferred parameters. The research results show that airlines should focus on Online boarding, In-flight Wi-Fi Service, Type of Travel and Class when improving customer satisfaction, which provides a reference for service-oriented companies to improve their economic efficiency.

Keywords: customer satisfaction, random forest, grid search

1. Introduction

Improving customer satisfaction can bring good reputation to the enterprise, which can not only improve the original customer's repurchase behavior but also attract more new consumers, thus making the enterprise invincible in the fierce market competition. For this reason, enterprises have to improve the quality of products or services, increase the value-added products or services, optimize the management process, etc., in order to maximize the satisfaction and even exceed customer expectations. Usually, customer satisfaction comes from consumers' personal experience and is real-time, so it requires enterprises to conduct customer satisfaction surveys and analyses from time to time, in order to accurately and timely understand consumer needs and improve products or services, so as to achieve continuous improvement of customer satisfaction. Since customer satisfaction plays such an important role in the survival and development of enterprises, both theoretical and practical circles have conducted in-depth research on it, covering a wide range of fields, including the connotation of customer satisfaction, customer satisfaction evaluation index, customer satisfaction factors, customer satisfaction prediction, customer satisfaction on business efficiency, etc. The research methods used include questionnaire survey method, structural equation method, multiple regression method, structural equation method, multiple regression analysis, literature analysis method, for example, Sheth & Howard [1] considered that customer satisfaction is a comparison between the price customers pay when they engage in consumption behavior and the benefits they receive after purchase. Oliver [2] suggested that customer satisfaction is a psychological state that consumers have when they engage in consumption behavior based on their own experience. Fornell

DOI: 10.54254/2754-1169/38/20231913

[3] stated that, Customer satisfaction is related to factors such as purchase price, customer expectations and customer perceptions, and can be directly assessed by comparing the perception of expectations before and after the consumption behaviour.Westbrook [4] suggests that customer satisfaction is a state of mind that occurs after a purchase has been made, and that customers are more likely to give positive feedback to products and businesses when their pre-purchase expectations are low and their expectations of the price of the product are not high. Domestically, Liu Baofa and Zou Zhaoju [5] reviewed the research results of customer satisfaction analysis model and concluded that the existing research results are more about the judgment of the customer's product satisfaction state at a certain point in time, which is a static model; however, the customer's satisfaction with a certain product is changing, which is a dynamic balance. Wu Yajuan [6] constructed an employment satisfaction evaluation system and 18 evaluation indicators based on Customer Satisfaction Index (CSI), and then used factor analysis and cluster analysis to evaluate and analyze each influencing factor and rank its importance. Hui Liu [7] analyzed the factors influencing the satisfaction of higher education students in China and constructed a satisfaction index model and satisfaction measurement index system based on the customer satisfaction index model established by western scholars.

Potential consumers usually refer to previous consumer reviews of a product or service to reduce the uncertainty they feel when purchasing it [8]. Park and Lee [9] found that the greater the intensity of online reviews had a greater impact on consumers' purchase intentions, and that negative reviews had a significantly greater impact on consumers' purchase intentions. The influence of negative reviews on consumers' purchase intention is significantly greater than the influence of positive reviews. With the development of computer technology and the application of big data, machine learning and other artificial intelligence methods are widely used by scholars for the analysis and prediction of customer satisfaction, among which the random forest algorithm is most commonly used. Based on big data mining technology, Liu Yang [10] used attribute selection method based on information gain rate to filter out user satisfaction related attributes from data, and used random forest algorithm to construct satisfaction evaluation prediction model, and finally used multi-tag classification algorithm to optimize the classification model. saumya [11] et al. used random forest classifier to classify reviews into low quality and high quality, and then used gradient augmented regressor for rating prediction. Baswardono [12] et al. demonstrate that the random forest algorithm has higher accuracy than the C4.5 algorithm for prediction by classifying airline customer satisfaction. Liu, Fan [13] conducted a study on customer satisfaction in the telecommunication industry and obtained a prediction model for telecommunication user satisfaction by integrating each business process based on the prediction results of logistic regression. Beibei Zhang and Min Hu [14] made improvements to the random forest algorithm using the grid search algorithm, and found that the improved random forest model was made to predict customer satisfaction accurately and efficiently, and then this model was used to rank the importance of satisfaction influencing factors, refine customer concerns, and extend customer satisfaction improvement strategies. Finally, the applicability and generalization of the optimization model is demonstrated by using an e-commerce platform as an example. The results of the study show that the word-of-mouth evaluation of merchants and the price level of meals for two are the main factors affecting customer satisfaction, and therefore, the restaurant industry should be given significant consideration when improving customer satisfaction.

In summary, it can be found that existing studies mainly collect data from consumer behavior records, questionnaires, online evaluations, etc., and apply logistic regression, random forest, neural network and text mining methods to carry out analysis and prediction research on the influencing factors of customer satisfaction, but most of the studies ignore the importance of parameter selection, and usually select parameters based on experience, this paper uses grid search for the random forest algorithm to In this paper, we use grid search to optimize the parameters of the random forest

DOI: 10.54254/2754-1169/38/20231913

algorithm, such as the number of decision trees *k* and the number of candidate splitting attributes *m* in the random forest, and analyze them with specific cases, in order to improve the accuracy and efficiency of predicting customer satisfaction scores.

2. Model building

2.1. Random Forest Algorithm

The random forest algorithm is an integrated decision tree learning algorithm, which is essentially a process of deriving multiple weak model decision trees for heavy model feature selection, regression problems, and classification problems. Due to the randomness of sample extraction and the randomness of feature attribute selection by the random forest algorithm, the error value fluctuates somewhat. In order to reduce the impact of its uncertainty on parameter selection and prediction results, this paper uses a grid search algorithm to optimize the random forest in order to take the optimal values of hyperparameters.

2.2. Confusion matrix

The confusion matrix is one of the original bases for the evaluation of machine learning models and is the basis for the derivation of a series of subsequent evaluation metrics. The column indexes of the confusion matrix are the true positive cases and the true negative cases, and the row indexes are the predicted negative cases and the predicted positive cases, respectively. When the true case is positive and the predicted case is also positive, the predicted sample is named True Positive, and the corresponding position in the confusion matrix (i.e., the position in the lower right corner of the matrix) is filled with the statistical value of this type of sample; similarly, there are also False Negative in the confusion matrix, i.e., the position in the lower left corner of the matrix; False Positive), which is the position in the upper right corner of the matrix; and True Negative, which is the position in the upper left corner of the matrix.

The sklearn tool provides mature implementations for all kinds of algorithmic models and the training, prediction, and evaluation processes of the models, then below we will use the confusion_matrix function of the sklearn tool metrics package to print the confusion matrix alone.

2.3. Model parameter calibration method

For each model of machine learning, some specific parameters (i.e., hyperparameters) need to be manually selected and set before the model is trained, and the parameter selection of the model largely affects the effectiveness of the model application. In order to reduce the impact of training bias caused by sampling randomness, we will use the Grid Search CV function of the sklearn tool model_seleetion library to perform hyperparameter tuning.

The grid search method is a straightforward, all-encompassing search strategy for parameter optimization. The essence of this method is to partition the parameter space into a number of grids, then optimize the model that will be trained by iteratively traversing all possible parameter combinations at the intersection of the grids while computing the root mean square error, also known as the accuracy, of the related model. The parameter combination that yields the maximum accuracy or the smallest root mean square error—i.e., the ideal parameter combination for the model—can only be determined by touring each node in the grid plane. In order to find the optimal solution of the model more expeditiously this experiment uses the grid search method to improve the random forest algorithm, using the model's least mean square error as the target of the optimal parameters, eliminating the influence of uncertainty largely is bypassed. on the training results helps prevent model misfit or overfit and ensures that the resulting solution is the best combination of parameters

DOI: 10.54254/2754-1169/38/20231913

over the entire mesh domain, avoiding significant problems that can arise when choosing traditional empirical adjustment methods. The solution obtained is the best combination of parameters in the full range of the lattice, avoiding the large errors that can be introduced by the traditional empirical method of parameter selection. Finally, referring to the work of Beibei Zhang and Min Hu [14], we will apply a grid search algorithm to tune and optimize two of the most important parameters in the Random Forest algorithm: the number of decision trees k and the number of possible ones partition attributes m.

3. Cases and Analysis

3.1. Data sources and pre-processing

The data for this study comes from the Airline Customer Satisfaction Dataset on the Kaggle platform, which provides user satisfaction ratings for over 120,000 passengers, including additional information on each passenger, flight and travel type as well as their rating of various factors such as cleanliness, comfort, service and overall experience. The dataset contains a large number of consumer and passenger reviews, making it an ideal online review source for predicting customer satisfaction on the Web with universal applicability.

There are 23 characteristic attributes in this dataset, and the factors with strong correlation from the analysis of the heat matrix of correlation coefficient are: Departure delay, Arrival delay, Cleanliness, Inflight entertainment, Ease of online booking and Online boarding, etc. The preliminary analysis shows that the factors with strong correlation with customer satisfaction include Customer Type, Type of Travel and Class. The specific analyses are shown in Figures 1, 2 and 3:

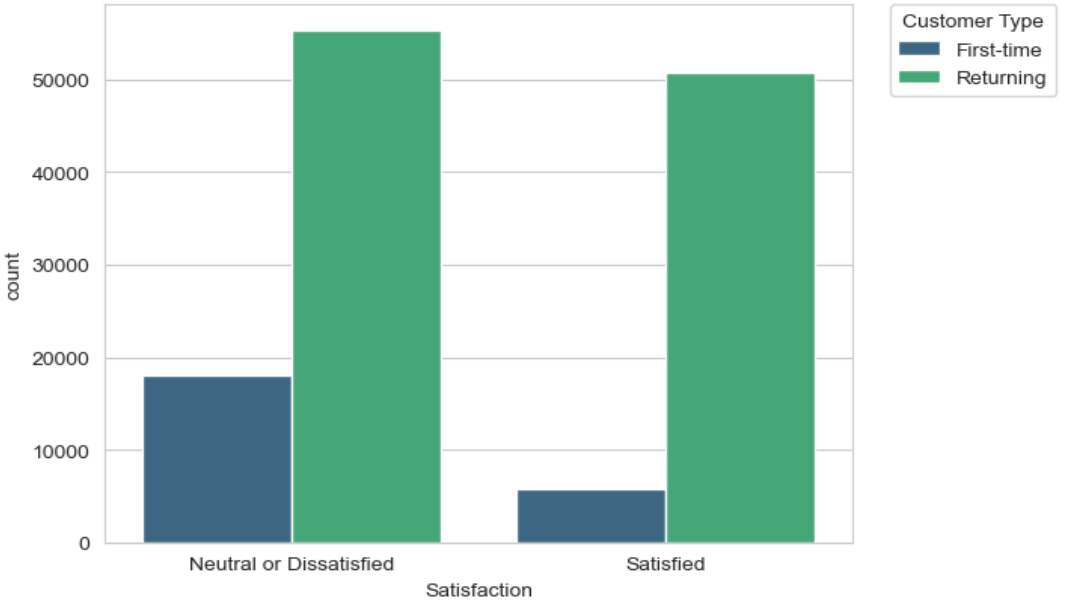

Figure 1: Relationship between customer type and customer satisfaction.

DOI: 10.54254/2754-1169/38/20231913

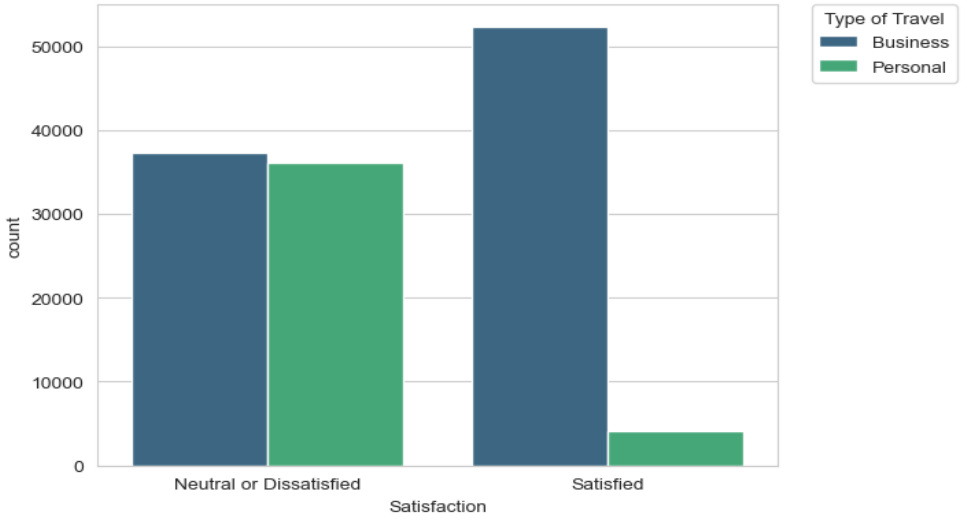

Figure 2: Relationship between trip type and customer satisfaction.

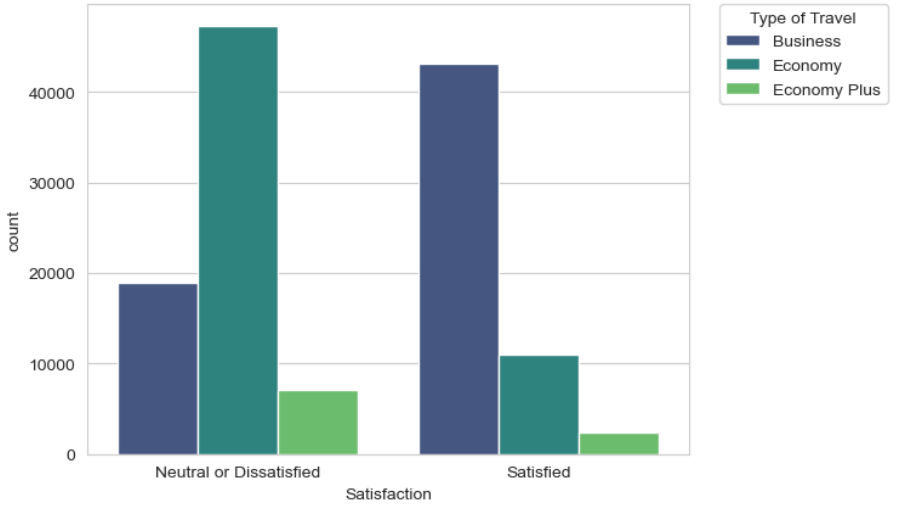

Figure 3: Relationship between position level and satisfaction.

In order to raise the efficiency of the prediction process and the accuracy of the results, data pre-processing is performed first. Firstly, we check the missing values and find that "Arrival Delay" is missing, so we use the average value to fill in the missing values; then we encode the classification data with one-hot and replace the original discrete data to raise the accuracy of the prediction results; then we prepare the training data, determine satisfaction as the target variable, extract 70% as the training Finally, StandardSealer is used to normalize and normalize the pre-processed data.

3.2. Forecast results and analysis

In this paper, the grid search method with five-fold cross-validation was used to adjust the random forest model with appropriate hyperparameters, and it was found that the optimized model had an even lower root mean square error of 0.0019 than the random forest prediction without hyperparameter configuration, and the prediction accuracy was improved. The values of each parameter of the grid search optimized random forest model are shown in Table 1.

DOI: 10.54254/2754-1169/38/20231913

Table 1: Parameter values of the grid search optimized random forest model.

Parameters	Data Type	Meaning	Parameter values
bootstrap	boolean	Whether data is extracted with or without release	Ture
max features	Int, float, stringy	Maximum number of features in the training set	[15,20,40]→40
n_estimators	integer	Number of decision trees	[30,40,50]→50
min_samples_ leaf	Int, float	Minimum number of samples of leaf nodes	1
min_samples_split	Int, float	Minimum number of samples required to delineate the nodes	2

Next, after dividing the training set, the random forest machine learning model is trained (trained with the optimal model hyperparameters searched by the grid search algorithm), and the prediction results are evaluated by performing dichotomous prediction on the test set for the completed training model. The evaluation index results and confusion matrix are shown in Fig. 4 and Fig. 5.

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.98	0.97	21937
1	0.97	0.94	0.95	17027
accuracy			0.96	38964
macro avg	0.96	0.96	0.96	38964
weighted avg	0.96	0.96	0.96	38964

Figure 4: Random forest model evaluation results.

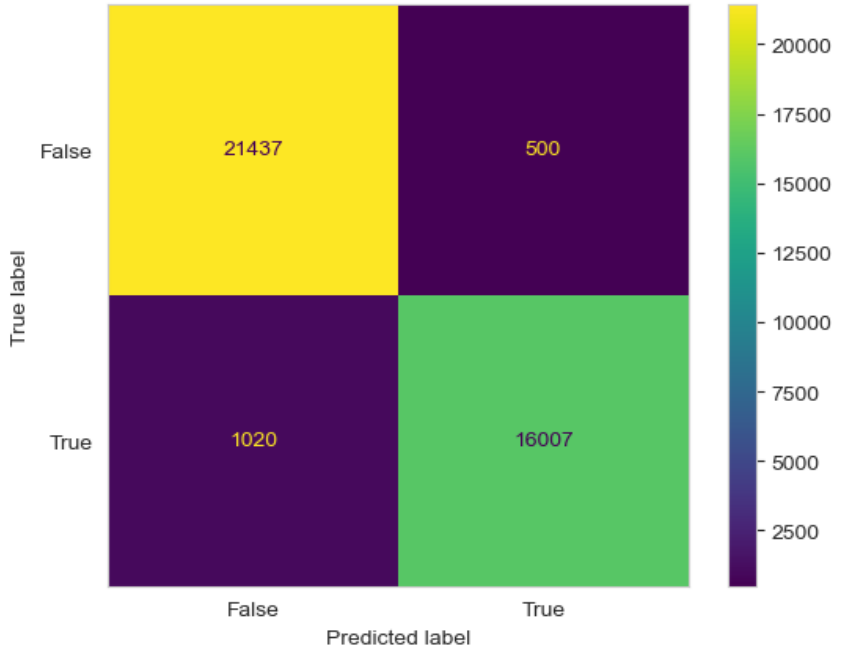

Figure 5: Confusion matrix of random forest model.

DOI: 10.54254/2754-1169/38/20231913

Finally, using the optimized prediction model, the importance of the input features is ranked, and the details of the importance ranking of the relevant customer satisfaction features are shown in Figure 6.

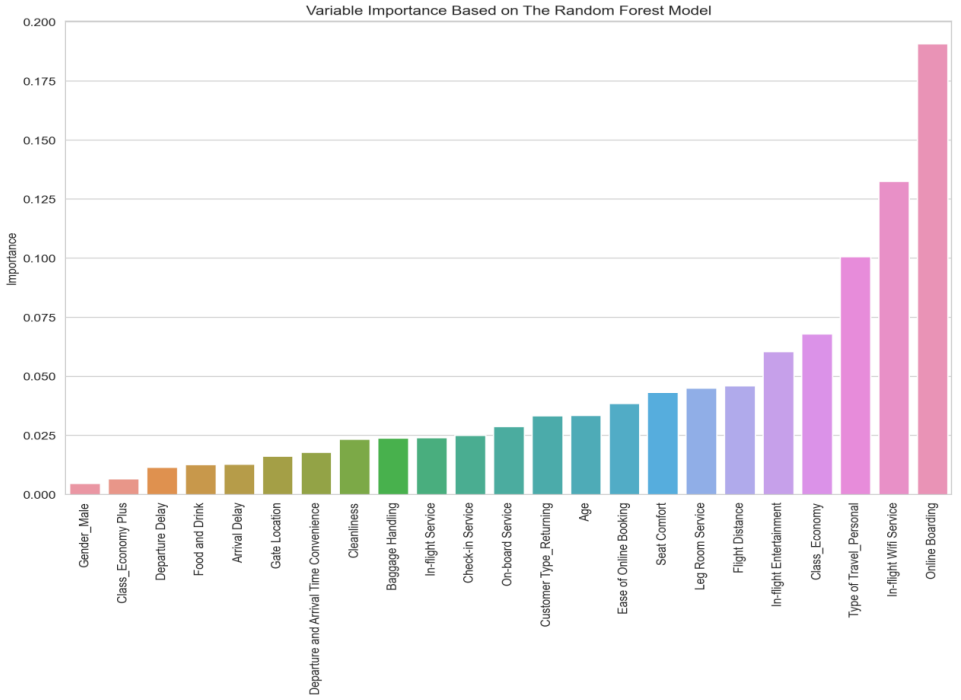

Figure 6: Ranking the importance of customer satisfaction features.

Figure 6 shows that the most important feature in the final prediction is Online boarding, followed by In-flight Wifi Service, and then Type of Travel_Personal, Class_Economy, In-flight Entertainment, Flight Distance, Leg Room Service, Seat Comfort, and Ease of Online Booking, while the other characteristics are less relevant to predicting customer satisfaction scores. Therefore, Online boarding, In-flight Wifi Service, Type of Travel and Class are important influencing factors of customer satisfaction. On the one hand, airlines should improve online services so that passengers can enjoy online check-in and in-flight wifi services conveniently to improve customer satisfaction; on the other hand, airlines can launch various preferential activities to reduce the cost of passengers to attract more passengers to travel by air.

3.3. Stability test

To ensure the reliability and validity of the prediction results, we trained three more machine learning models, logistic regression, KNN and support vector machine (trained with the optimal model hyperparameters searched by the grid search algorithm), and performed dichotomous prediction on the test set for the completed models to evaluate the prediction results. The results of evaluation metrics and confusion matrix are shown below.

DOI: 10.54254/2754-1169/38/20231913

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.90	0.89	21937
1	0.87	0.84	0.85	17027
accuracy			0.87	38964
macro avg	0.87	0.87	0.87	38964
weighted avg	0.87	0.87	0.87	38964

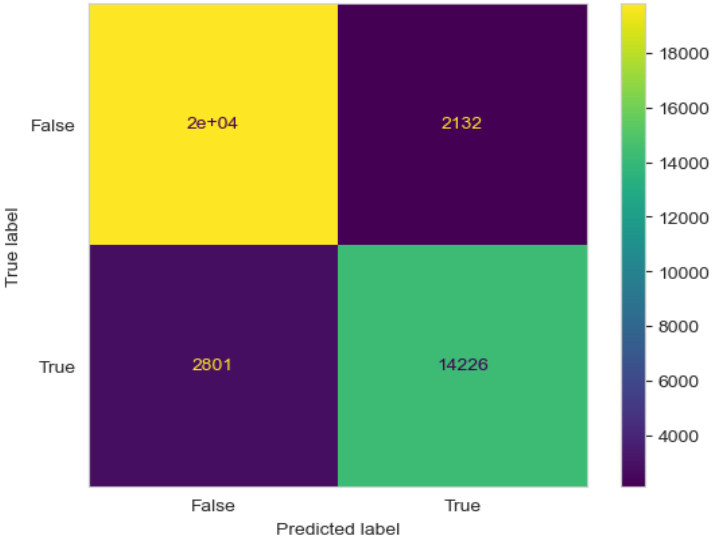

Figure 7: Logistic regression model evaluation results with confusion matrix.

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.96	0.94	21937
1	0.94	0.89	0.91	17027
accuracy			0.93	38964
macro avg	0.93	0.92	0.93	38964
weighted avg	0.93	0.93	0.93	38964

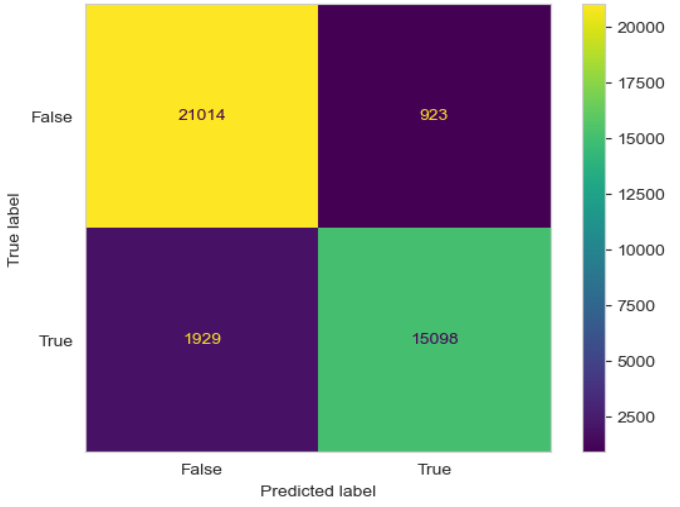

Figure 8: KNN model evaluation results with confusion matrix.

DOI: 10.54254/2754-1169/38/20231913

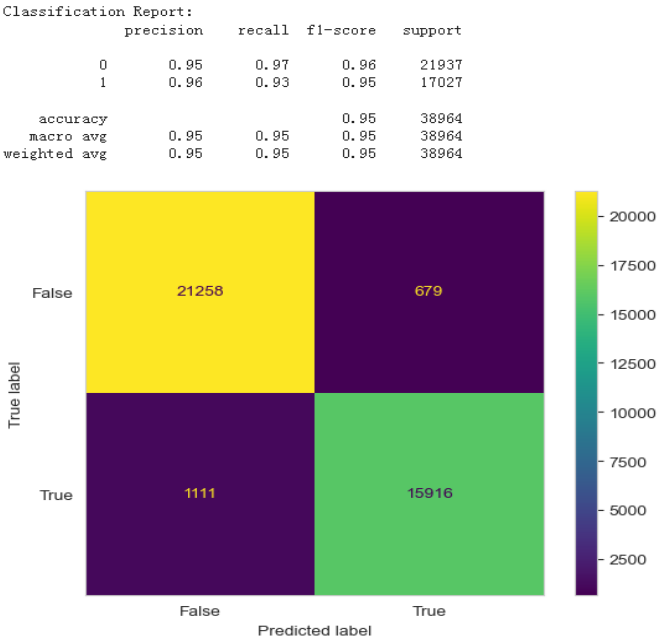

Figure 9: Support vector machine model evaluation results with confusion matrix.

From Figures 7, 8, and 9, it can be found that, the four machine learning models trained on customer satisfaction prediction results evaluate basically the same results, and the random forest model evaluates the best results.

4. Conclusion

In order to quickly respond to customer-driven market changes, it is of great importance to study customer satisfaction prediction. In this paper, the number of decision trees and the number of candidates split feature attributes of random forest are optimally tuned by using grid search algorithm to improve the shortcomings of random forest, which can effectively predict the customer satisfaction scores of airlines. The model can be applied to the example of customer satisfaction prediction and can be extended to the satisfaction prediction of other service companies. Among the examples used in this study, Online boarding, In-flight Wi-Fi Service, Type of Travel and Class are important factors influencing customer satisfaction, so airlines should consider them when improving customer satisfaction.

References

[1] *Sheth J N, Howard J A. The Theory o} Buyer Behavior [M]. New York:Wiley, 1969.*
[2] *Oliver R.L, Linda G. Effect of Satisfaction and Its Antecedents on Consumer Preference and Intention [J].Advances in Consumer Research 1981,8(1):88- 93.*
[3] *Fornell G, Liu J L, K.ang J. A national customer satisfaction barometer: the swedish experience [J]. Journal of Marketing, 1992,56(1):6-22.*
[4] *Westbxook R A. Intrapersonal Affective Influences on Consumer Satisfaction With Products [J]. Journal of Consumes Research, 2003, 7(1):49-54.*
[5] *Liu, Baofa, Zou, Zhaoju. Judgment and prediction model of customer satisfaction[J]. Science and Technology Management Research, 2005, (2):202-204*
[6] *Wu Yajuan. Employment satisfaction statistics and prediction of college students based on factor one cluster analysis[J]. Journal of Nanjing University of Information Engineering (Natural Science Edition), 2010,2(06):510-513.*
[7] *Liu Hui. A study on measuring student satisfaction in Chinese higher education based on PLS-SEM [D]. Jiangsu University, 2011.*

DOI: 10.54254/2754-1169/38/20231913

[8] Chatterjee P. Online reviews: do consumers use them? [J]. Advances in Consumer Research, 2001,28(1):129-134.
[9] Cheol Park, T'hae Min Lee. information direction, website reputation and eWOM effect: A moderating role of product type [J]. Journal of Business Research, 2007,62(1):61-67.
[10] Liu Yang. User multidimensional satisfaction evaluation prediction system [D]. Nanjing University,2016.
[11] Sunil Saumya, Jyoti Prakash Singh, Abdullah Mohammed Baabdullah, et al. Ranking online consumer reviews [J]. Electronic Commerce Research and Applications,2018,29:78-89.
[12] Baswardono W, Kurniadi D, Mulyani A, et al. Comparative analysis of decision tree algorithms: Random Forest and C4.5 for airlines customer satisfaction classification [J]. Journal of Physics: Conference Series, 2019, 1402 (6):066055.
[13] Liu Fan. Research on telecommunication user satisfaction prediction[D]. Zhejiang University of Technology and Industry,2020.
[14] Zhang, Beibei, Hu, Min. Improved random forest based on grid search for customer satisfaction prediction[J]. Journal of Beijing University of Information Science and Technology, 2021,36(04):50-53+58.