Prediction of New York taxi tip behavior based on machine learning classification and regression methods

Hejingyu Huang[1,*]

[1]Geographic Information Science at Beijing Forestry University, Haidian District, Beijing, China

*Corresponding author: 18177262970@163.com

Abstract. In the context of machine learning, this study employs a learning method to process big data and predict and analyze taxi tip behavior. Basic variables such as trip time, trip distance, and number of passengers are added to the dataset, as well as special variables related to geographic location. Using these variables, a two-stage model is created in which a random forest classification model with an accuracy rate of 98.3% in the first stage and a Lasso regression model with an MSE value of 0.007294 in the second stage are used to predict taxi tip behavior, resulting in better fitting than a single model prediction.

keywords: machine learning, taxi prediction, classification, regression

1 Introduction

Taxi tips serve as assessments of driver services and play a crucial role in determining their income. Despite the widespread practice of tipping, the amount of gratuities received by drivers remains unpredictable, posing a challenge for both drivers and passengers. Traditional tip prediction methods usually rely on manual and statistical analysis [1], but these methods are prone to limitations and errors. In recent years, machine learning methods [2][3] have made significant advances in their ability to build predictive models by utilizing historical data and feature extraction. However, for the fundamental task of tip behavior prediction, a single classification or regression model may not achieve the highest level of precision. This paper attempts to find a reasonable method to predict tip behavior, providing more accurate and useful information for drivers, passengers, etc., and improving the efficiency and experience of travel. Through research, this paper finally divides the problem into two parts and completes the study through classification and regression models as a two-stage model, respectively predicting the likelihood of New York passengers providing tips as well as the percentage of tips given by passengers in relation to the total cost of the trip [4]. The classification model [5][6] uses the random forest model, with an accuracy of 98.3%, while the regression model uses the Lasso model [7], with an MSE of 0.007365. This generated model offers valuable insights for taxi drivers in New York,

Z. Wang et al. (eds.), *Proceedings of the 2023 2nd International Conference on Public Service, Economic Management and Sustainable Development (PESD 2023)*, Advances in Economics, Business and Management Research 273,
https://doi.org/10.2991/978-94-6463-344-3_75

including selecting areas where passengers are more inclined to give more tips, and for taxi companies to arrange operation time and location. In addition, it can also provide guidance for passengers on whether to give tips and how much to give.

2 Research Methods

2.1 Data introduction

The New York taxi data used in this study was obtained from the NYC Taxi & Limousine Commission. The dataset used contains a total of 271,519 records, each representing a trip taken by a driver in New York City between January 1, 2013, 0:00, and January 31, 2013, 24:00.

2.1.1 Dataset Used in the Study.

This study focuses on the behavior of passengers giving tips, so the percentage of tips in the total fare (tip/tot) and the label of whether a tip was given are the main variables of interest (y). As for other variables (x), this study mainly focuses on variables such as the time of pickup, the number of passengers, and the duration of the trip. The distribution of categorical and numerical variables in the dataset is shown in Table 1.

Table 1. Distribution of Variable Types.

Type	Number	Percentage
Categorical	11	57.89%
Numerical	8	42.11%

2.1.2 Variable Description.

Table 2. Meanings of numerical variables.

Names	Types	Meanings
rate_code	Numerical	The final rate code applied at the end of the trip.
trip_time_in_secs	Numerical	The duration of the trip in seconds.
trip_distance	Numerical	The distance of the trip in miles.
fare_amount	Numerical	The base fare of the trip.
surcharge	Numerical	Additional charges.
tolls_amount	Numerical	Tolls paid during the trip.
count	Numerical	The number of pickups by the driver during the time period.
percent	Numerical	The percentage of pickups where the driver received a tip during the time period.

Table 3. Meanings of categorical variables

Names	Types	Meanings	variable differentiation
hack_license	Categorical	The driver's license number.	N/A
passenger_count	Categorical	The number of passengers.	N/A
payment_type	Categorical	The code representing how the passenger paid for the trip.	N/A
weekday	Categorical	The weekday when the driver made the pickup.	N/A
pickup_hour	Categorical	The hour when the driver made the pickup.	N/A
sd_label	Categorical	The distance between the pickup location and Times Square.	200m
her_label	Categorical	The distance between the pickup location and Wall Street.	200m
CU_label	Categorical	The distance between the pickup location and Columbia University.	200m
NYU_label	Categorical	The distance between the pickup location and New York University.	200m
KA_label	Categorical	The distance between the pickup location and Kennedy International Airport.	1000m
LA_label	Categorical	The distance between the pickup location and LaGuardia Airport.	1000m

Explanation: (1) As shown in Table 3,since airport coordinates are a single value rather than a range, and airports cover a large geographic area, the values chosen to differentiate the KA_label and LA_label variables as labels are 1000m. (2) As shown in Table 2,passenger_count, weekday, and pickup_hour are selected as numerical categorical variables because the number of passengers, day of the week, and hour do not have a linear effect on the problem, and therefore treating them as categorical variables is more reasonable. (3) Moreover, considering that the geographic location at the time of departure may also have a certain impact on tip behavior, this paper incorporates six landmark geographic locations in New York(As shown in Table 3). By studying these variables, some differences between landmark locations in New York City can be reflected, providing insight into the differences between regions and different types of landmarks (such as restaurants, squares, schools, airports). (4) The distances between two points computed from the geographic variables mentioned above, as well as the distances used to distinguish whether a trip started from a certain location, are all straight-line distances.

2.2 Problem Handling

Model Selection: Before the start of the study, the author only used regression models to try to predict tips. However, since 48.55% of the data has a tip value of 0, directly performing regression analysis on the data with many 0 values will result in lower

accuracy. Therefore, after practical comparisons, a two-stage model is chosen, which uses a classification model to predict whether a passenger will give a tip first, and then uses a regression model to predict how much tip a passenger might give if they do give one. This approach is more reasonable from both a practical and mathematical perspective.

2.3 Classification Problem

2.3.1 Data Preprocessing.

Step 1: First, search the entire data set to check for missing and zero values in each column. Based on the query results, this paper uses pairwise deletion in direct deletion method to handle missing values, only deleting cases with missing values in the required research variables [8]. The column "store_and_fwd_flag" is deleted due to its high missing value rate of 80.413%, and 5641 rows with missing longitude and latitude, accounting for approximately 2.078% of the data, are also removed.

Step 2: Process the outliers in the variables caused by system errors or manual errors. (1) Remove trips with a duration of less than 120 seconds. (2) Remove trips with a distance of less than 0.2 miles.

Step 3: Manipulate the relevant columns in the data to obtain the required variables according to mathematical relationships. (1) Count the number of occurrences of each driver's license number as the driver's experience "count." (2) Connect each driver's license number with the labels to calculate the probability of receiving a tip "percent." (3) The six variables related to geographic locations are all new categorical variables calculated by mathematical formulas(as shown below) based on the distance between the pickup location and special locations (sd_label, her_label, CU_label, NYU_label, KA_label, LA_label). Formula [9]:

$$\text{Distance} = 6371004^{*}\ \text{ACOS}\ ((\text{SIN}(\text{RADIANS}(X1))\ ^{*}\ \text{SIN}\ (\text{RADIANS}(X3)) +\text{COS}(\text{RADIANS}(X1))^{*}\text{COS}(\text{RADIANS}(X3))^{*}\text{COS}(\text{RADIANS}(X2\text{-}X4))) \quad (1)$$

$$X1=\textit{starting\_latitude},\ X2=\textit{starting\_longitude},\ X3=\textit{special\_location\_latitude},\ X4=\textit{special\_location\_longitude} \quad (2)$$

2.3.2 The Impact of Various Variables on Tipping.

Due to the large number of values in some variables, they are divided into 10 equal-frequency bins for plotting. The y-axis represents the proportion of passengers who gave tips among all passengers. This study uses many variables, and only a few variables that show certain patterns are presented here for observation.

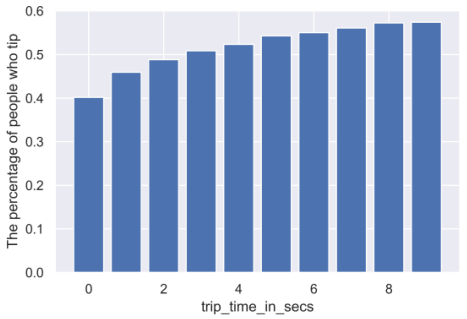

Fig. 1. The impact of time on tipping.

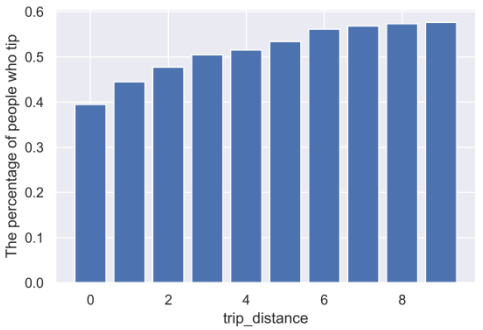

Fig. 2. The impact of distance on tipping.

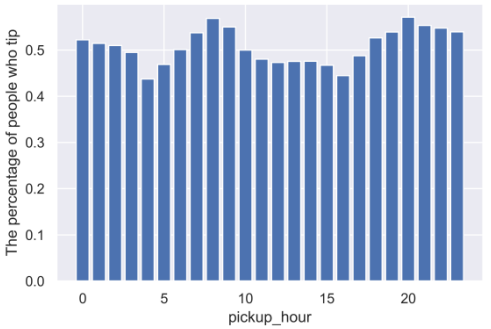

Fig. 3. The impact of hour on tipping.

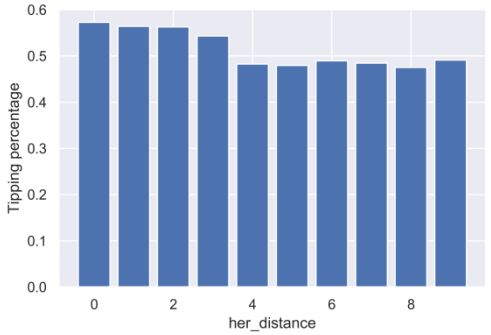

Fig. 4. The impact of from the starting point to Wall Street on tipping.

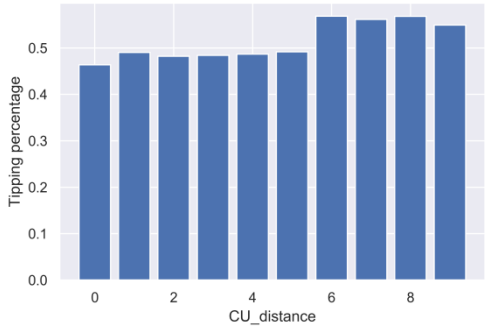

Fig. 5. The impact of distance from the starting point to Columbia Uni. on tipping.

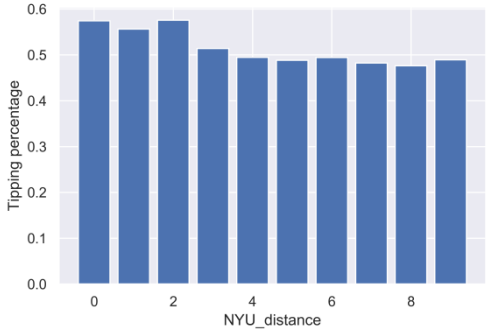

Fig. 6. The impact of distance from the starting point to NY Uni. on tipping.

(1) Time: As trip duration increases, passengers tend to give more tips in recognition of the additional time and effort required by the driver, leading to a greater tendency for passengers to give tips. (Figure 1) (2) Distance: As distance increases, passengers tend to give more tips for the same reason as time, as longer trips require

more time and effort from the driver. (Figure 2) (3) Hour: The highest likelihood of tipping occurs during rush hour at 8, 9, 20, and 21. In contrast, the likelihood of tipping is lower during early mornings, afternoons, and evenings, specifically at 4, 5, 6, 14, 15, and 16. In the highest tipping time frame, tipping is more likely when passengers arrive on time during rush hour. In the lowest tipping time frame, there are fewer people taking taxis during this time, and taking a taxi may not be for the purpose of going to work on time, leading to a lower likelihood of tipping. (Figure 3) (4) Wall Street: The closer the distance to Wall Street, the higher the likelihood of tipping, and as the distance increases, the likelihood gradually decreases and tends to stabilize. (Figure 4) (5) Columbia University: In contrast to Wall Street, the closer the distance to Columbia University, the lower the likelihood of tipping, and as the distance increases, the likelihood of tipping gradually increases. (Figure 5) (6) New York University: Unlike Columbia University, the closer the distance to New York University, the higher the likelihood of tipping, and as the distance increases, the likelihood gradually decreases. (Figure 6)

2.3.3 Transforming Skewed Continuous Features and Normalization.

After examining each variable, two variables are found to be highly skewed. To improve the accuracy of the study, logarithmic transformation is applied to these two variables, which effectively reduces the range of abnormal data. It should be noted that since the logarithm of 0 is meaningless, the "tolls_amount" variable should be treated with x+1.

In the numerical variables studied in this paper, each variable has a different dimension. To increase the accuracy of classification processing, normalization is applied to make the dimensions of each variable the same.

2.3.4 Hyperparameter Tuning and Prediction.

Step 1: Split all data into a training set and a testing set in a 0.7:0.3 ratio, with a random seed of 0. After the split, the training set contains 186002 data points, including 19 variables x and 1 variable y, while the testing set contains 79716 data points, including 19 variables x and 1 variable y.

Step 2: After processing the numerical variables, one-hot encoding is applied to the remaining 11 categorical variables to convert them from characters to numbers for classification analysis. The "get_dummies" function is used in this paper, resulting in a total of 560 columns of data.

Step 3: The grid search method is used for hyperparameter tuning. First, logistic regression is employed for tuning, adjusting C and penalty, which takes 241.856 seconds. Second, decision tree is used for tuning, adjusting max_depth and min_samples_split, which takes 310.632 seconds. Third, random forest is used for tuning, adjusting max_depth and n_estimators, which takes 1319.599 seconds. After comparison, the random forest tuning effect is more ideal, so that it should be adopted as the classification model for this paper, with an accuracy of 0.983350.

The effect of the random forest model obtained is shown in the Table 4 below:

Table 4. Random Forest Model Classification Effect

	Accuracy	Precision	Recall	F1	Auc
Training Set	0.983484	0.999908	0.966101	0.982714	0.991541
Testing Set	0.983115	0.999678	0.965434	0.982258	0.983619

2.4 Regression Problem

Building on the classification problem completed in the previous section, a regression model is used to predict how much tip the passenger would give.

2.4.1 Data Processing.

Initially, the paper conducted regression analysis on all of the data. However, after preprocessing, hyperparameter tuning, and prediction analysis, the accuracy was not deemed to be optimal. Through comparative studies, it was found that this might be due to the large proportion of 0 values in the data, leading to prediction deviations and lower accuracy. Therefore, to eliminate the influence of too many 0 values on the regression prediction, the paper first removes the 0 values in the tip before fitting the model, and only trains and predicts with non-zero data.

2.4.2 Tip Ratio Chart.

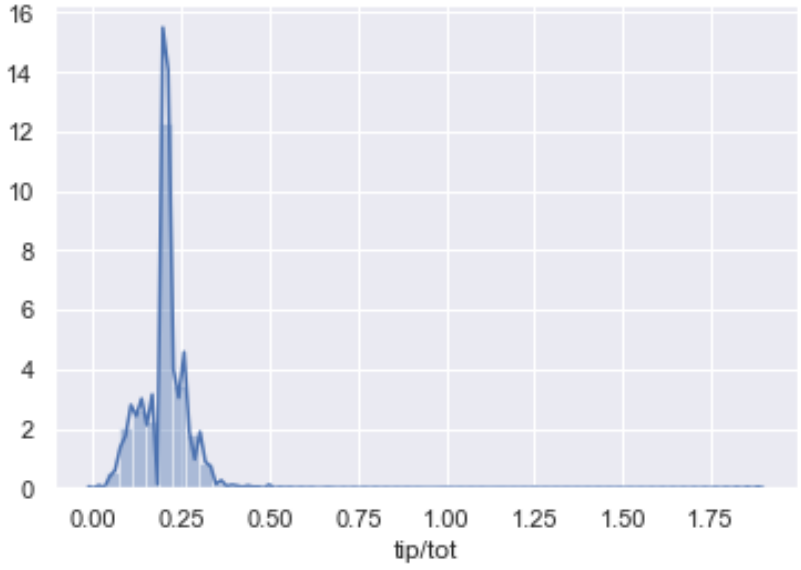

Fig. 7. Tip distribution less than 200%.

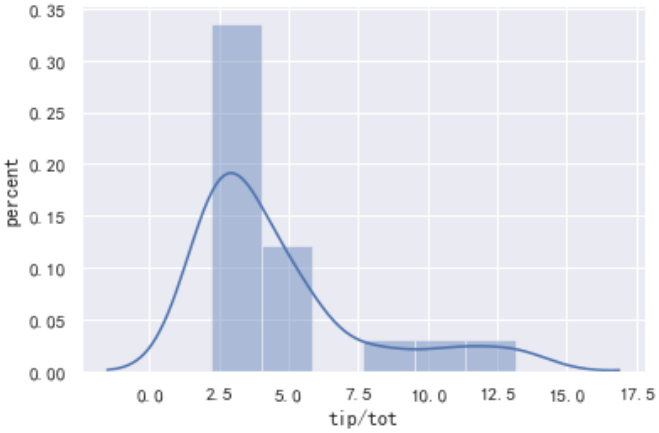

Fig. 8. Tip distribution greater than 200%.

Considering that tipping is a variable with a significant subjective component, it is possible that wealthy passengers may provide higher tips. Furthermore, instances of extremely large tips are quite uncommon. Therefore, the paper chooses not to exclude those high tips from the analysis. To avoid the small proportion of tips not being clearly represented in the chart, the paper draws separate charts for tips ratios less than or equal to 2 and greater than or equal to 2 to observe the distribution of tip ratios.As shown in Figure 7 and Figure 8.

2.4.2 Tip Ratio Chart.

(1) As time increases, the percentage of tip gradually decreases. This may be because most passengers give a fixed tip, and as the total fare increases, the percentage of tip decreases. (Figure 9) (2) As the distance increases, the percentage of tip given by passengers gradually decreases. A longer trip may lead to a poor experience for passengers, and the driver may take a longer route, resulting in less tip given by passengers. (Figure 10) (3) The highest percentage of tip given is during the time periods of 0-5am and 5-7pm. The reason for this may be that there are fewer passengers during 0-5am, and the drivers work harder during this time, resulting in higher tip given by passengers. The reason for the peak during 5-7pm may be due to rush hour or dinner time, with more passengers and a higher percentage of tip given during these periods. (Figure 11) (4) As the distance between the origin of the trip and Times Square increases, a slight negative correlation is observed in the percentage of tip bestowed upon the driver by passengers. The influx of individuals in close proximity to Times Square is notable, and those who visit or consume there are likely to have more financial resources. Therefore, the closer the distance to Times Square, the higher the likelihood of the driver receiving a more substantial gratuity. (Figure 12)

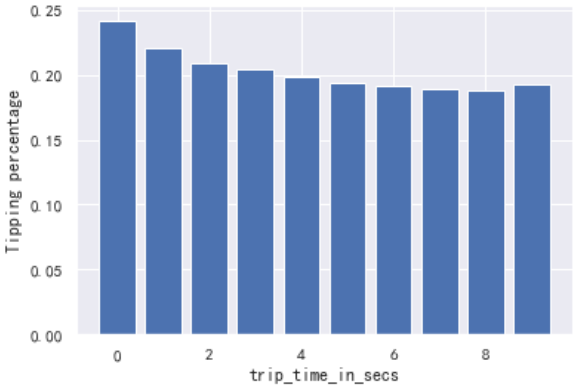

Fig. 9. The impact of time on tipping.

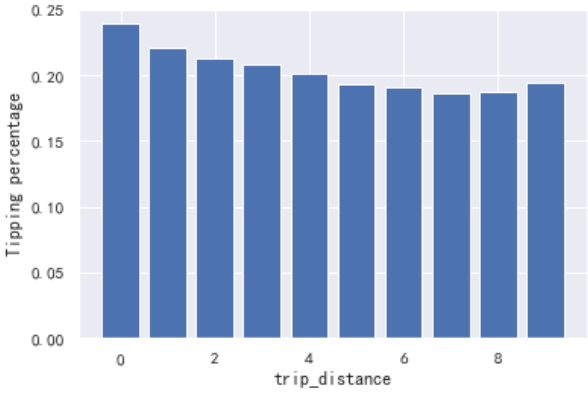

Fig. 10. The impact of distance on tipping.

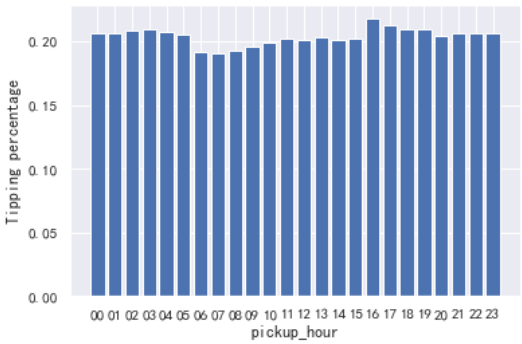

Fig. 11. The impact of hour on tipping.

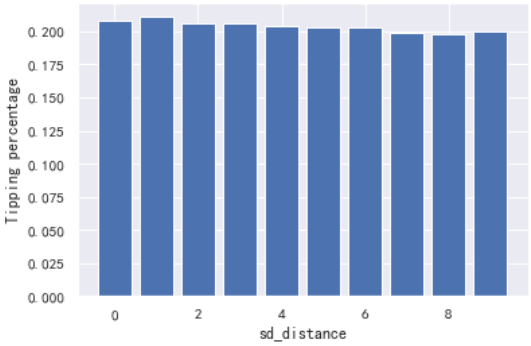

Fig. 12. The impact of from the starting point to Time Square on tipping.

2.4.3 One-Hot Encoding.

Similar to the classification problem, categorical variables are encoded using one-hot encoding techniques in the pursuit of solving regression problems. Specifically, the "get_dummies" function is employed to convert such variables into a format that is amenable to processing, resulting in the creation of a total of 558 data columns.

2.4.4 Hyperparameter Tuning and Prediction.

Step 1: Split all data into a training set and a testing set in a 0.7:0.3 ratio, with a random seed of 0. After the split, the training set contains 95718 data points, including 558 variables x and 1 variable y, while the testing set contains 41022 data points, including 558 variables x and 1 variable y.

Step 2: Hyperparameter tuning: Similar to the classification model, the grid search method is adopted, and the linear regression model is first used for fitting. The Lasso regression model is used for tuning in the second round, adjusting its alpha value. The regression results are shown in Table 5 below:

The testing set MSE for the linear regression model is 0.007365, and the testing set MSE for the Lasso regression model is 0.007294. Therefore, after comparison, the Lasso regression model is chosen as the final model for the regression problem.

Table 5. Regression Results

Lasso	mse	mae
Training Set	0.009612	0.045022
Testing Set	0.007294	0.044099

3 Conclusion

This study aims to develop a two-stage classification and regression model for the purpose of forecasting passengers' tipping behavior in New York City and to enhance the accuracy of such predictions. Through the classification phase, a remarkable pre-

diction accuracy rate of 98.3% is achieved, thus illustrating the effectiveness of this method in tackling analogous problems. While this research has achieved promising results, there remain some challenges that require further investigation.

Firstly, the parameter tuning time in this study is too long, with a random forest model taking up to 1300 seconds for tuning. In order to further improve model performance, future research can explore methods to shorten the parameter tuning time to find the best model parameters more quickly.

Secondly, future research can explore more and more appropriate variables, and add relevant geographical location variables to optimize the model. These variables can further improve the predictive ability of the model, and better explain the correlation between tipping behavior and geographical location.

Finally, model selection and parameter tuning in the regression problem in this study can still be adjusted to make the final results more accurate. Future research can explore more regression models to further optimize model performance.

In summary, this study shows that the two-stage model constructed using classification and regression has high accuracy, and can effectively predict the tipping behavior of passengers in New York City, providing guidance for drivers, passengers, and even taxi companies. Future research can explore more methods to optimize the model to improve its accuracy and better explain tipping behavior.

References

1. Y. Kim, J. Lee. Predicting taxi drivers' tip earnings using GPS data. In Proceedings of the 2010 ITSC (pp. 176-181). IEEE., (2010).
2. P. Zhen, A brief discussion on machine learning methods [J]. NSTA, No.157(01): 176-177. (2014).
3. M. l. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects. Sci. 349,255-260. (2015).
4. N. Jin, Analysis of hotel occupancy rate prediction based on tourism data [D]. Yanshan University, (2018).
5. L. Breiman, Random Forests for Regression and Classification. Machine Learning, 45(1), 5-32. (2001).
6. A. Liaw, M. Wiener, Classification and regression by random Forest. R News, 2(3), 18-22. (2002).
7. S. He, G. Zhao, Q. Cui, Research on the influencing factors of school evaluation results based on Lasso-logistic regression and random forest models [J]. Journal of Changchun Normal University, 41(02): 11-16. (2022).
8. X. Yang, Comparison of missing data imputation methods from a predictive perspective [D]. Southwest University of Finance and Economics. (2021).
9. J. Zhao, L. Liu, Z. Wu, Calculation of distance between two microwave stations and azimuth angle of microwave antenna based on longitude and latitude [J]. Inner Mongolia Radio and Television Technology,29(04): 49-50+53. (2012).