Tech Science Press

Journal of Quantum Computing
DOI:10.32604/jqc.2021.016315
Article

Flight Delay Prediction Using Gradient Boosting Machine Learning Classifiers

Mingdao Lu, Peng Wei, Mingshu He[*] and Yinglei Teng

Beijing University of Posts and Telecommunications, Beijing, 100876, China
[*]Corresponding Author: Mingshu He. Email: hemingshu@bupt.edu.cn

Received: 16 February 2021; Accepted: 13 April 2021

Abstract: With the increasing of civil aviation business, flight delay has become a key problem in civil aviation field in recent years, which has brought a considerable economic impact to airlines and related industries. The delay prediction of specific flights is very important for airlines' plan, airport resource allocation, insurance company strategy and personal arrangement. The influence factors of flight delay have high complexity and non-linear relationship. The different situations of various regions and airports, and even the deviation of airport or airline arrangement all have certain influence on flight delay, which makes the prediction more difficult. In view of the limitations of the existing delay prediction models, this paper proposes a flight delay prediction model with more generalization ability and corresponding machine learning classification algorithm. This model fully exploits temporal and spatial characteristics of higher dimensions, such as the influence of preceding flights, the situation of departure and landing airports, and the overall situation of flights on the same route. In the process of machine learning, the model is trained with historical data and tested with the latest actual data. The test result shows that the model and this machine learning algorithm can provide an effective method for the prediction of flight delay.

Keywords: Delay prediction; machine learning; gradient boosting

1 Introduction

With the rapid development of China's civil aviation industry, the impact of flight delays has become increasingly prominent. According to the Statistical Bulletin on the Development of the Civil Aviation Industry of China in 2019, the cumulative passenger transport volume in 2019 was 65,993,420,000 people, an increase of 7.9% over the previous year. 49,662,000 sorties were completed in the whole year, and 1.352 billion passenger throughputs were completed at civil aviation airports. But the frequency and severity of flight delay is not optimistic, the average normal rate of all flights in the country is only 81.65%.

The causes of flight delays are currently more difficult to explain due to multiple and repetitive factors, such as weather, airport takeoff or landing management, airline management, air traffic, air traffic control, passenger reasons, and so on [1]. For airports, flight delays will result in the disruption of limited airport resource allocation arrangements such as limited routes, runways, aprons and so on, which will increase the pressure on Airport security, operation and resource scheduling. For airlines, flight delays will result in increased operating, maintenance and human costs, which will seriously affect costs and profits. For passengers, flight delays cause irreparable losses to personal travel arrangements or business travel. For insurance companies, flight delay prediction is very important for the pricing and operation of travel insurance and flight delay insurance. The classification predication of flight delays in this paper will help to improve the above problems.

The generalization ability of flight delay prediction model and the multidimensional and complex factors that affect flight delay make flight delay prediction a challenging task [2]. How to construct effective features from original data plays an important role in machine learning system. Other factors, such as weather, delays in upstream flights, the conditions of departure and landing airport are dynamic and play an important role in highlighting feature selection in accurate predictions.

This paper conducts model training, validation and testing using flight information of different airports and routes in China from 2016 to 2017. The data processing is detailed in Part 5. Through the analysis of specific business scenarios, we abstract weather impact, departure and landing airport management and resource impact factors as quantitative features by establishing high-dimensional features that are in line with business reality, thus further improving the accuracy of delay prediction. Machine learning uses gradient-lifting XGBoost and GBDT methods and compares them with traditional machine learning classification models. Finally, the actual data of 59105 different airports and routes sampled in October 2020 are used to test the multi-classification prediction model, which achieves 88.11% multi-classification accuracy with 5 minutes, 15 minutes and 30 minutes delay of planned landing time, and has good portability and generalization ability.

This paper has innovations in the following aspects:

1. Using gradient boosting machine learning algorithm with more extensive training data to construct a prediction algorithm with stronger generalization ability, which is not limited to specific airports or routes;

2. Combined with the actual business scenario analysis, the high-dimensional training features are abstracted from the airport resources and management, aviation management information and weather factors. In addition to the time dimension, the feature engineering adds space dimension of air route analysis;

3. In this paper, the multi classification delay prediction is carried out, and has achieved good result. The effect of delay regression analysis is not satisfactory, and the classification prediction of delay is more practical in many business scenarios.

2 Related Work

In the past decade, domestic and foreign scholars mainly focus on the analysis of flight delay factors, the establishment of delay propagation model and the solutions to alleviate the delay.

In the field of flight delay prediction, according to the existing literature research methods can be roughly divided into three categories, inferential prediction model based on statistical theory, prediction method based on simulation and empirical model, and delay prediction model based on machine learning. This section will briefly describe the first two methods, and focuse on the current situation of prediction research based on machine learning method.

Based on the statistical theory, an inferential prediction model can be established. Based on the actual historical sample data, the characteristics of the sample data are analyzed by statistical theory, thus the statistical model is continuously fitted, approximated and optimized. The commonly used methods and models are regression analysis, Bayesian network, Caiman filtering, etc. Li et al. [3] analyzed the complex factors affecting the delay, and established a Bayesian network model based on statistical methods to predict the delay of downstream flights.

The prediction method based on simulation and empirical model is used to simulate aircraft operation model, flight delay propagation model and airport management scheduling model. The key variables are connected with the whole simulation model system to predict the flight delay time in each simulation scenario.

The delay prediction method based on machine learning, through the data processing of a large number of actual historical samples, extracting key features, and then put the results into the machine learning algorithm for model establishment. And finally, through the model to achieve flight delay prediction. This method needs a lot of data to fit the model. When the correlation between the factors is complex or nonlinear, it will be better to use machine learning method to study. The machine learning method commonly used in prediction research is supervised learning.

In the process of machine learning, a prediction model is established based on a set of marked training data. Through the feature vector obtained from a large number of data input, the regression or classifier parameters are continuously fitted to make the model achieve the expected performance. It is the most common learning method to solve the regression or classification problems. The common algorithms include artificial neural network, support vector machine, decision tree, etc. Fig. 1 shows a training process of supervised learning model.

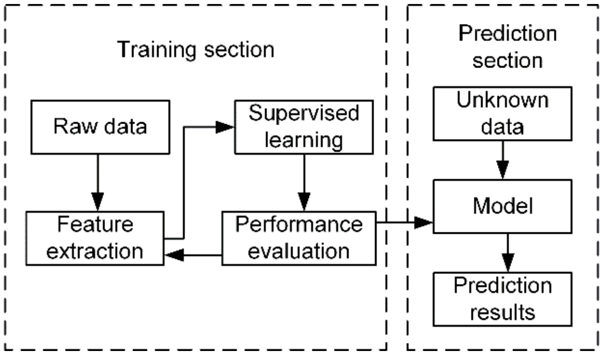

Figure 1: Training process of supervised learning model

Engin et al. [4] used artificial neural network to predict flight delay directly. Only a specific airport was selected, and the prediction accuracy was about 90% when aircraft type information was considered. Reboll et al. [5] used the random forest algorithm to predict the delay between 2 and 24 h, established the air traffic network model and took the delay state as the influence variable of delay. When the delay threshold was set at 60 min, the average error of 100 network sections was 19%. Navoneel et al. [6] used common machine learning classifiers, random forest, gradient lifting classifier and support vector machine to predict flight delay of several large airports in the United States. The accuracy of delay prediction was 79%. Choi et al. [7] also used decision tree, random forest and other algorithms with adding weather factors to the feature engineering for prediction, and the accuracy rate of random forest was 80%. Etani [8] used the classification model of random forest, added weather data and flight data as training features, and made correlation analysis. The prediction accuracy for punctual arrival of flights reached 77%.

3 Model Description

3.1 Algorithm Introduction

In this paper, the prediction model is essentially a classification task. Due to the complexity of mapping between features and prediction tags, ensemble learning method is proposed to train and construct the model. Ensemble learning is to construct multiple classifiers (weak classifiers) to predict the data set, and then use a certain strategy to integrate the prediction results of multiple classifiers as the final prediction results [9]. In this paper, boosting serial mode is used for training base classifier. There is a certain dependence between each base classifier. Its basic idea is to stack the base classifiers layer by layer. When each layer is trained, it will give a higher weight to the wrong samples of the previous base classifier. When testing, the final result is obtained by weighting the results of each classifier.

Different from bagging classifier, the latter has no strong dependence among base classifiers and can be trained in parallel. The prediction result is decided by all parallel classifiers, such as random forest algorithm.

The base classifier uses cart regression tree model to complete the construction of the whole tree by continuously splitting the features [10]. For example, the current tree node is split based on the j-th eigenvalue. Let the sample whose eigenvalue is less than s be divided into left subtree, and the sample larger than s is divided into right subtree. Formula expression is shown in Eq. (1).

$$R_1(j,s)=\{x|x^{(j)}\leq s\} \text{ and } R_2(j,s)= \{x|x^{(j)}>s\} \tag{1}$$

Its essence is to divide the sample space on the feature dimension, which is very complex. For the objective function $L=\sum_{x_i \in R_m}\left(y_i\text{-}f(x_i)\right)^2$ of each cart tree, the best segmentation feature and the optimal segmentation point s are solved, i.e., Eq. (2), then a base classifier is finally determined.

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i\text{-}c_1)^2+\min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i\text{-}c_2)^2\right] \tag{2}$$

Next, for the integration relationship of the base classifiers, we can use the concatenation method, each time let the next weak classifier to fit the residual error of the error function to the predicted value (the residual is the error between the predicted value and the real value) [11]. For example, in GBDT (gradient boosting decision tree) algorithm, when we define the prediction loss function of each weak classifier as the mean square error function $l\left(y_i,y^i\right)=\frac{1}{2}\left(y_i\text{-}y^i\right)^2$. It will calculate the negative gradient result as $\text{-}\left[\frac{\partial l\left(y_i,y^i\right)}{\partial y^i}\right]=\left(y_i\text{-}y^i\right)$. Therefore, when the mean square loss function is selected as the loss function, the value of each fitting is (Real value- Value predicted by the current model), i.e., residual. In this case, the variable is "the value of the current prediction model". That is to say, we have to calculated the negative gradient.

Finally, the scores of multiple weak classifiers are accumulated to get the final prediction, and each iteration is based on the existing ensemble model, adding a tree to fit the residual between the prediction results of the integrated model and the real value, which has a good effect on the prediction classification.

Like the traditional boosting tree model, XGBoost algorithm also adopts residual learning method to improve the model. The difference is that the selection of split nodes of weak classifier is not necessarily based on the least square loss. Similarly, when we want to predict the score of a sample, according to the features of the sample, it will fall on the corresponding leaf node in each tree, and each leaf node corresponds to a score. Finally, the score corresponding to each tree is added up to get the predicted value of the sample [12]. The expression is shown in Eq. (3), $w_{q(x)}$ is the score of leaf node q, and f(x) is the expression of one of the regression trees.

$$\hat{y}=\phi(x_i)=\sum_{k=1}^{M}f_k(x_i)$$

$$\text{where } F=\left\{f(x)=w_{q(x)}\right\}\left(q:R^m \rightarrow T, w \in R^{T}_{/\!/ n}\right) \tag{3}$$

Xgboost objective function is defined as $L^{(t)}=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t\text{-}1)}+f_t(x_i)\right)+\Omega(f_t)$. The first part is used to measure the difference between predicted score and true score. The second part is regular term, which can control the number of leaf nodes and node scores to prevent over fitting. Next, we will determine the f_t function that minimizes the objective function. The idea of XGBoost is to approximate it with its Taylor second-order expansion at $f_t = 0$. So the objective function is approximated to $L^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i,\hat{y}^{(t\text{-}1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)$, where g_i is the first derivative and h_i is the second derivative. Because the residual of the prediction score of the first T-1 tree has no effect on the optimization of the objective function, the first part can be directly removed. At the same time, each sample ultimately falls in the leaf node, and the samples of the same leaf node can be combined. The simplification formula is as Eq. (4).

$$L^{(t)} \simeq \sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)$$

$$= \sum_{i=1}^{n}\left[g_i w_q(x_i)+\frac{1}{2}h_i w_{q(x_i)}^2\right]+\gamma T+\lambda\frac{1}{2}\sum_{j=1}^{T} w_j^2 \tag{4}$$

For the fixed structure, the optimal weight of leaf node can be calculated, see Eq. (5). And then the objective function can be obtained. After the nodes and weights are determined, the next tree can be fitted or the fitting process is finished.

$$w_j^{*}=\text{-}\frac{G_j}{H_j+\lambda} \quad L=\text{-}\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_i+\lambda}+\gamma T_B \tag{5}$$

Both GBDT and XGBoost algorithms are fitted by boosting integration, and continue to learn through residual error, which have good effects in classification tasks.

3.2 Parameter Adjustment

Since both GBDT and XGBoost use cart regression decision tree, the algorithm parameters are basically derived from the decision tree. In order to prevent over fitting or under fitting state, in this paper, n_estimators, learning_rate, max_depth and other important parameters are brought into the selection model, and the GridSearchCV function grid search method in scikit-learn package is used. In this method, the parameters are adjusted according to the step size within the specified parameter range, and the parameters with the highest accuracy in the verification set can be found from all the parameters [13]. The results of best parameters with grid search are shown in the Tab. 1 below.

Table 1: The results of best parameters with grid search

Hyper-parameters	Value	Hyper-parameters	Value
‘learning_rate’	0.09	‘n_estimators’	500
‘max_depth’	11	‘min_samples_split’	9
‘min_samples_leaf’	10	‘min_samples_split’	14

GridSearchCV function can ensure that the parameters with the highest accuracy can be found within the specified parameter range [14], but this is also the defect of grid search. This method requires traversing the combination of all possible parameters, which is time-consuming in the case of large data sets and multi parameters.

3.3 An Overview of Feature Selection of Models

When the parameters of the model and algorithm are determined, we need to select a feature subset from all the features, so that the constructed model has better effect and stronger generalization ability. Selecting an optimal subset from all features or constructing more abstract features can make the best performance in training and testing data for certain evaluation criteria. The detailed feature selection and feature extraction process is in the fourth and fifth parts.

4 Data Processing and Feature Establishment

Before applying the algorithm to our dataset, we need to perform a basic preprocessing. Data preprocessing is to transform the data into a format suitable for our analysis, and also to improve the quality of data. The obtained data sets are incomplete, noisy and inconsistent. This paper obtains the data sets of Chinese flights from 2015 to 2017 from the public flight information. The dataset consists of 19 columns and 328291 rows. The original data of flight information is shown in the Tab. 2.

Table 2: The original data of flight information

Airport of departure	Airport of landing	Flight number	Planned departure time (×10^9)	Planned arrival time (×10^9)	Actual takeoff time (×10^9)	Actual arrival time (×10^9)	Aircraft number
AKU	URC	CZ6919	1.4518656	1.4518704	1.451865	1.451869	1
HRB	NKG	AQ1040	1.4541333	1.4541438	1.454134	1.454144	2
NKG	KMG	MU2719	1.4525550	1.4525661	1.452556	1.452565	3
HGH	AVA	GJ8733	1.4529846	1.4529936	1.452985	1.452995	4
KHN	TNA	JD5145	1.4529975	1.4530038	1.452998	1.453003	5

Firstly, the missing data value is processed. The key information (such as flight departure and landing time) in many rows of data is missing or empty, which cannot be estimated or approximately filled, and can only be discarded. Use the dropna() function in the pandas package to clean up the dataset and delete rows and columns from null values that contain key information. After preprocessing, the key attributes of 1716 data are null, and the number of rows is reduced to 326575.

Then the time data is processed. Machine learning algorithm can't learn the feature of second, which is a serious interference to the prediction model. According to the planned take-off and arrival time, the planned flight time is calculated as a feature, which can indirectly reflect the route distance [15]. The planned departure and arrival dates are calculated to match the local weather and airport information. The flight month, departure time and arrival time are calculated as the features of model. The arrival delay time is calculated as the target value of delay and the delay time of the preceding flight.

Then we extract more features. According to the flight delay propagation model, the influence of preceding flights on downstream flights is the most direct embodiment. By screening the same aircraft number, the aircraft is sorted by time. If the arrival airport of the upstream flight on the same day is the departure airport of the downstream flight, then we consider that the two flights are adjacent flights. For the flight data with later time, the feature of preceding flight delay is filled by the arrival delay time of the earlier flight data. In this paper, the mean filling method is used for the flights with missing preamble delay.

According to the analysis of actual business scenarios, in order to characterize the impact of departure and landing airport and flight delay propagation on the same route, three more abstract features are extracted: the overall average delay of the same route, the average delay of the same route 2 h before the target flight landing, and the average delay of landing airport 2 h before the target flight landing. The specific analysis of the three characteristics is in the fifth part. Using the grouby() function to group the preprocessed information, judge the scheduled arrival time of flight, and calculate the average delay of flights within 2 h. Extract flight information and flight nature according to flight number. The first two digits of the flight number are the airline information of the flight. The flight with the last letter of the flight number is a supplementary flight. The flight number has three digits for domestic flights and four digits for international flights.

Finally, the local weather and airport information match with flight information. Code the special weather and mark the flights with special events in the planned time.

The general flow chart of data processing and feature establishment is shown in Fig. 2 below.

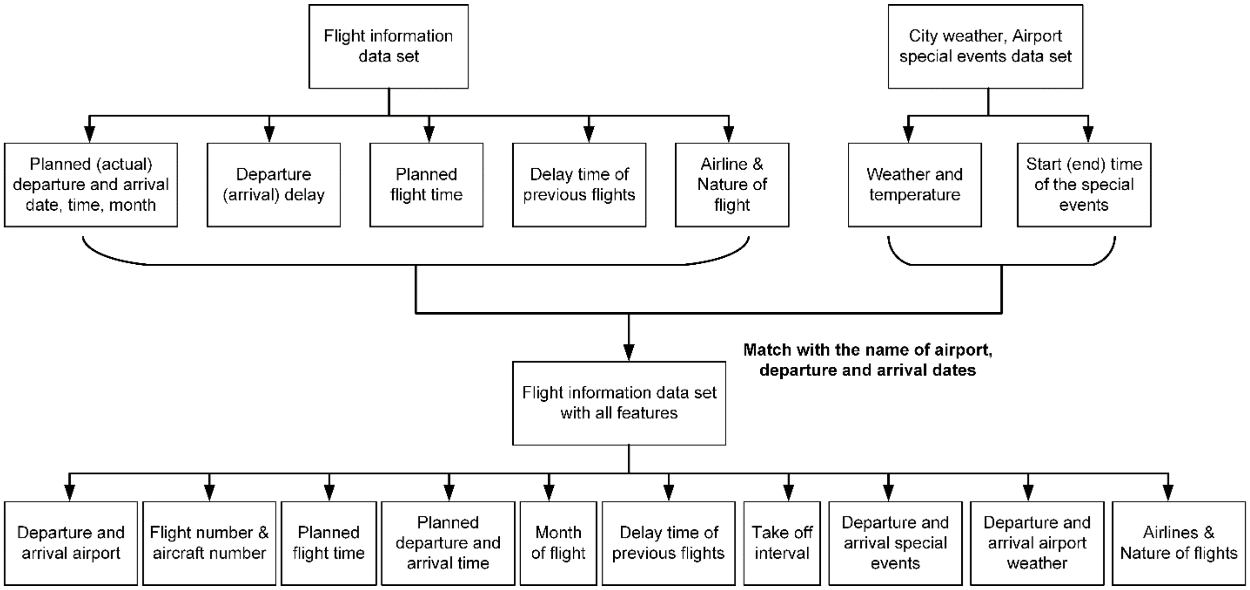
Flight information data set
City weather, Airport special events data set
Planned (actual) departure and arrival date, time, month
Departure (arrival) delay
Planned flight time
Delay time of previous flights
Airline & Nature of flight
Weather and temperature
Start (end) time of the special events
Match with the name of airport, departure and arrival dates
Flight information data set with all features
Departure and arrival airport
Flight number & aircraft number
Planned flight time
Planned departure and arrival time
Month of flight
Delay time of previous flights
Take off interval
Departure and arrival special events
Departure and arrival airport weather
Airlines & Nature of flights

Figure 2: Flow chart of data processing and feature establishment

5 Feature Selection and Extraction Combined with Business Analysis

When using machine learning algorithm to build the model, it is to fit the learning model continuously through the characteristics of samples and corresponding labels. The importance of feature selection is selfevident. When the sample data features are less, we should even consider adding features. In fact, many features are often redundant. In this section, in addition to using the conventional method to filter features, the real business scenarios are analyzed, which abstracts the features difficult to quantified into higher dimensional statistical features, and tries to add spatial features.

5.1 Feature Selection of Embedded Method

After using gradient boosting algorithm, we use embedded method for feature selection. The weight coefficients of each feature are obtained by the trained model, and the filter selection is carried out according to the coefficient weight. Of course, chi square test and maximum information coefficient (MIC) can also be used to evaluate the feature parameters, so as to filter the parameters with low correlation. Rutuja et al. [16] analyzed the factors influencing flight delay and the degree of influence on flight delay through the "selecting K best" method test, and obtained the most significant 12 features.

Embedding method is to carry out feature selection process and algorithm training at the same time. Firstly, the algorithm and model of machine learning are used for training to learn which features contribute the most to the accuracy of the model [17]. The weight coefficient of each feature represents the importance of the feature to the model, so as to find the most useful feature for the model accuracy. Compared with the filtering method, the embedding method has a more direct and better effect on improving the effectiveness of the model. The process of feature selection is shown in Fig. 3.

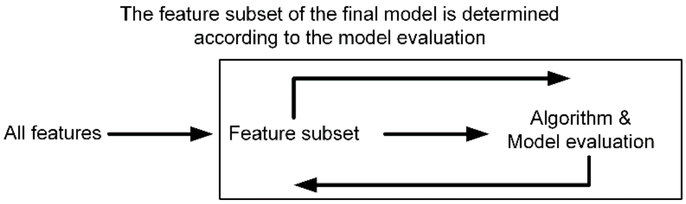
The feature subset of the final model is determined
according to the model evaluation
All features
Feature subset
Algorithm &
Model evaluation

Figure 3: The process of feature selection

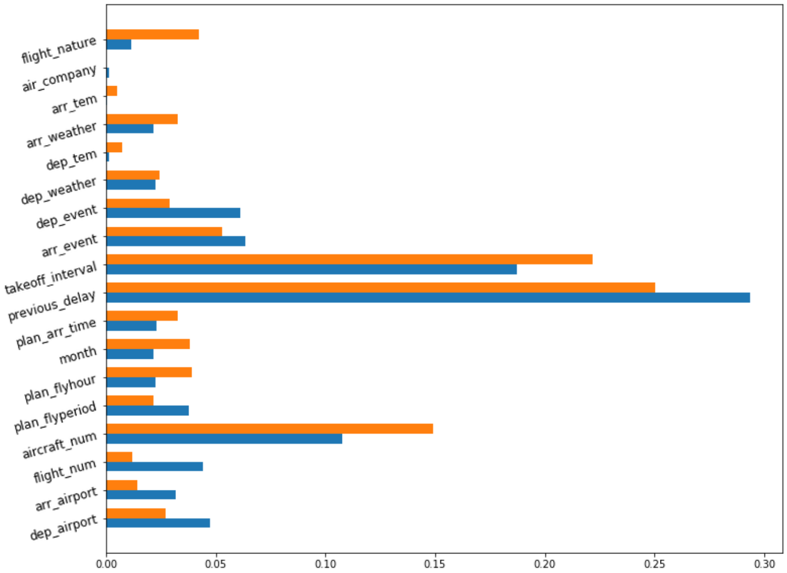

Figure 4: The importance distribution of the initial feature set

From the fourth part of data processing and feature establishment, 18 preliminary features are extracted and generated according to the original data. When the number of features is large and all kinds of features contribute to the model, it is difficult to define the effective critical screening value by filtering method. Using the method "feature_ importances" in scikit-learn package to score and evaluate the features, the importance of feature contribution to the model is calculated. Fig. 4 shows the importance distribution of the initial feature set. The blue histogram is the feature importance distribution of GBDT model, and orange is the feature importance distribution of XGBoost model.

From the analysis of the importance of features, it can be seen that under the two algorithms, some features contribute little to the prediction model, such as airline company, take-off and landing temperature and weather. At the same time, the original data set is further analyzed. The special circumstances data of take-off and landing airports are very incomplete and easy to have a great impact on the prediction. The two features are screened out at the same time for the features with low correlation, we select them out and put others back into the model for training and fitting. Fig. 5 shows the distribution of importance after feature selection. Tab. 3 shows the prediction accuracy of the model on the test set using different feature sets. The blue histogram is the feature importance distribution of GBDT model, and orange is the feature importance distribution of XGBoost model. According to Tab. 3, GBDT algorithm and XGBoost algorithm have obvious improvement in accuracy after feature selection.

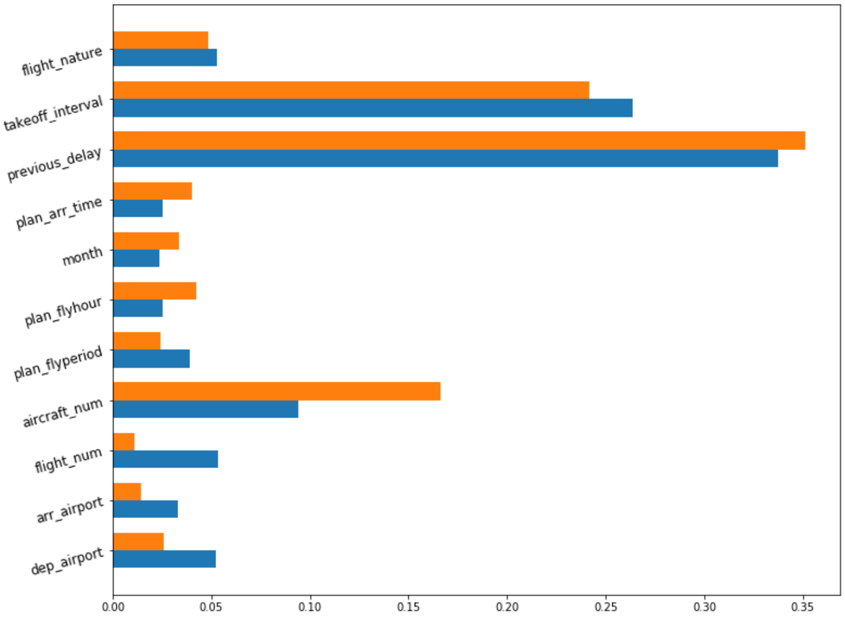

Figure 5: The distribution of importance after feature selection

Table 3: Comparison of model accuracy

Algorithm	Before feature selection		After feature selection	
	Accuracy of training set	Accuracy of test set	Accuracy of training set	Accuracy of test set
GBDT	80.58%	78.52%	83.12%	80.44%
XGBoost	78.09%	78.06%	79.34%	79.91%

5.2 Feature Extraction with Business Analysis

In the process of fitting model with machine learning algorithm, the processing of data and features is the basis of the whole process. Processing the original data and extracting effective features will achieve better results. In many cases, even if the algorithm selection or parameters are not optimal, it can still achieve good results. In this paper, the time attribute is divided into multiple dimensions, such as year, month, day, and hour of takeoff and landing, so as to avoid misleading the model through learning the trend of delay by seconds. At the same time, the classification attributes such as departure airport, arrival airport, departure weather, arrival weather, flight number, departure temperature, arrival temperature and airlines are coded.

According to the actual situation of business analysis, the impact of bad weather of departure and landing airport is the main reason for flight delay. However, the weather factors for takeoff and landing are mainly visibility, low-level cloud conditions, thunderstorm area distribution, strong crosswind and so on, as well as the high-altitude weather factors in the flight route. These factors are not directly obtained from local weather data. Even in some high-altitude thunderstorm areas, due to the strict restriction of civil aviation lines, there is little room for detour and gyration, which will lead to flight delays. At the same time, we consider a series of follow-up conditions caused by bad weather, such as runway icing, ponding, etc. Such delay factors cannot be extracted directly from simple data [18].

Similarly, in the analysis of the delay caused by the allocation of airport resources, there is no fixed quantitative arrangement statistics for runway, apron, corridor bridge and basic support of flight transit. In the actual operation, manual decision has great influence. Yao et al. [19] established a flight delay propagation prediction model, considering the critical resources of the aircraft, crew and the airport. They proposed that the impact of airports or flight resources on flight delays is not a simple and fixed linear model. In extreme cases, the resources occupied by the upstream delayed flights may even lead to the delay of the three downstream flights, so a simple delay model cannot be used for resource utilization assessment. In the process of model training, we cannot simply construct the features by using the airport resources before the flight landing, otherwise it is easy to generate more noise for the training.

According to the above business analysis and reference, this paper obtains three higher latitude features based on a large amount of data analysis in feature engineering: Average flight delay of the same route (Feature A), Average delay of flights on the same route 2 h before landing (Feature B), and Average delay of flights at the same landing airport 2 h before landing (Feature C). For the three features, we use two different algorithms for training and testing, each with the same parameters. The comparison results are shown in Tab. 4.

Table 4: Comparison of prediction results of models with different new features

New features used	GBDT		XGBoost	
	Accuracy of training set	Accuracy of test set	Accuracy of training set	Accuracy of test set
No new features	83.12%	80.44%	79.34%	79.91%
Feature A	84.32%	80.96%	80.72%	80.60%
Feature B	84.49%	81.08%	80.80%	80.56%
Feature C	84.24%	80.53%	80.58%	81.23%
Feature A and Feature B	85.68%	81.92%	81.95%	81.36%
Feature A and Feature C	84.69%	80.98%	80.85%	80.54%
Feature B and Feature C	85.58%	81.51%	81.74%	81.14%
Feature A and Feature B and Feature C	86.74%	82.87%	82.81%	82.48%

After adding new features in Tab. 4, the accuracy of the two algorithms has been significantly improved. Among them, GBDT algorithm has high sensitivity to single addition of feature B, and the accuracy rate is improved by about 1%. When adding multiple features, adding three new features can improve the accuracy by about 2%. XGBoost model has the biggest improvement for single feature C. When three new features are added, the accuracy can be improved by about 2.5%.

6 Result Analysis

After preprocessing and feature extraction, 75% of the original data are selected for training model and 25% for test. From the fifth part, we can see that the accuracy of 82.87% and 82.48% is obtained on the test set of GBDT and XGBoost. In order to further verify the prediction and generalization capability of the algorithm and model, we used GBDT model to test the actual data of 295525 different airports and different routes sampled in October and November 2020. The prediction task is still a multi classification of 5 min, 15 min and 30 min delay with planned landing time.

The training accuracy of training set and test set are 88.28% and 88.11%, respectively, and the prediction confusion matrix of test set is shown in Tab. 5.

Table 5: The prediction confusion matrix

Confusion matrix		Prediction results			
		Category 1	Category 2	Category 3	Category 4
Actual data	Category 1	15172	1489	239	40
	Category 2	1811	19627	1166	76
	Category 3	307	1171	12671	331
	Category 4	45	57	296	4607

The multi classification evaluation indicators basically adopts the confusion matrix indicators of two categories, but it cannot be directly calculated in recall, precision and F1 score. For multi classification, these indicators can be divided into macro average and micro average. Macro average is to calculate all kinds of recall and precision indicators, and then average them to get the final value [20]. Micro average is to calculate the average values of TP, FP, TN and FN, and then calculate the overall recall and precision indicators. Here we use the macro average to calculate the evaluation indicators, as shown in Tab. 6.

Table 6: Evaluation indicators of classification

Evaluating indicators	Value
Precision	0.8106
Recall	0.8042
F1-score	0.8083

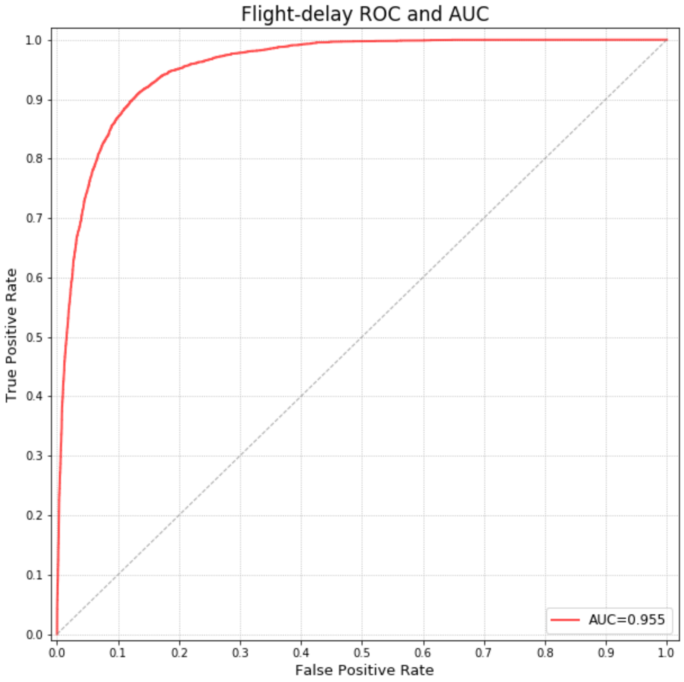

Figure 6: ROC Curve and AUC

In order to draw the ROC curve of multi classification, we recode the label of each test sample. The position of ‘1’ in the code indicates the sample category, and other positions are 0. If the classifier classifies the samples correctly, the value of the position corresponding to 1 in the probability matrix P is larger than that of the position corresponding to 0. Based on these two points, the label matrix L and the probability matrix P are expanded by row, and two columns are formed after transposition. This method can directly get the final ROC curve after calculation, as shown in Fig. 6. The calculated AUC value is 0.9886, which proves that the prediction classification method has high authenticity.

7 Conclusion

This paper proposes a flight delay prediction algorithm based on gradient boosting machine learning classifiers. It fully excavates the more abstract features of the influence of the preceding flight, the situation of departure and landing airport, the overall situation of the same route flight, and can analyze and forecast the flight delay. The results of series experiments show that the algorithm can predict flight delay by classification with high accuracy, and has good reliability in analysis and prediction.

The future research work includes the application of more advanced and appropriate preprocessing technology and artificial intelligence algorithm to obtain better performance. In this paper, we use the flight data covering all the airports in China, and have good generalization prediction ability after testing. Therefore, the model can also be trained with data from other countries in the future. More accurate forecasting models can be developed by using a mixture of complex models and many other models with appropriate processing capabilities, as well as using larger detailed data sets. The classification task can also be transformed into delay regression analysis, which can provide more accurate delay analysis and reference for airport resource arrangement, airline decision-making and personal travel.

Acknowledgement: Thanks for the experimental environment provided by laboratory ICN&CAD of School of Electronic Engineering, Beijing University of Posts and Telecommunications. And thank Mei Song, Xiaojuan Wang, Min Zhang, Xinlei Wang and Keqi Liu for their great help and company in this work.

Funding Statement: This work was supported by the National Natural Science Foundation of China (61871046).

Conflicts of Interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] X. Geng, "Analysis and countermeasures to flight delay based on statistical data," in *2013 5th Int. Conf. on Intelligent Human-Machine Systems and Cybernetics*, Hangzhou, China, pp. 535–537, 2013.

[2] P. Meel, M. Singhal, M. Tanwar and N. Saini, "Predicting flight delays with error calculation using machine learned classifiers," in *2020 7th Int. Conf. on Signal Processing and Integrated Networks*, Noida, India, pp. 71–76, 2020.

[3] Q. Y. Li, L. Wang, R. Fei, B. Wang, X. H. Hei, "An analysis method for flight delays based on Bayesian network," in *27th Chinese Control and Decision Conf.*, Qingdao, pp. 2561–2565, 2015.

[4] E. Demir and V. B. Demir, "Predicting flight delays with artificial neural networks: Case study of an airport," in *2017 25th Signal Processing and Communications Applications Conf.*, Antalya, pp. 1–4, 2017.

[5] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 231–241, 2014.

[6] N. Chakrabarty, "Flight arrival delay prediction using gradient boosting classifier," in *Emerging Technologies in Data Mining and Information Security,* 1st ed., vol. 1. Singapore, pp. 651–659, 2019.

[7] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conf.*, Sacramento, CA, pp. 1–6, 2016.

[8] N. Etani, "Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data," *Journal of Big Data*, vol. 6, no 1, pp. 85–87, 2019.

[9] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta *et al,* "A statistical approach to predict flight delay using gradient boosted decision tree," in *2017 Int. Conf. on Computational Intelligence in Data Science*, Chennai, pp. 1–5, 2017.

[10] J. Ye, J. H. Chow, J. Chen and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in *18th ACM Conf. on Information and Knowledge Management*, pp. 2061–2064, 2009.

[11] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[12] T. Hastie, R. Tibshirani and J. Friedman, "Boosting and additive trees," in *The Elements of Statistical Learning*, Springer, pp. 337–387, 2009.

[13] N. Chakrabarty, "A data mining approach to flight arrival delay prediction for American airlines," in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conf.*, Jaipur, India, pp. 102–107, 2019.

[14] F. Liu, J. Sun, M. Liu, J. Yang and G. Gui, "Generalized flight delay prediction method using gradient boosting decision tree," in *2020 IEEE 91st Vehicular Technology Conf.*, Antwerp, Belgium, pp. 1–5, 2020.

[15] L. Moreira, C. Dantas, L. Oliveira, J. Soares and E. Ogasawara, "On evaluating data preprocessing methods for machine learning models for flight delays," in *Int. Joint Conf. on Neural Networks*, pp. 1–8, 2018.

[16] R. Dhanawade, M. Deo, N. Khanna and R. V. Deolekar, "Analyzing factors influencing flight delay prediction," in *2019 6th Int. Conf. on Computing for Sustainable Global Development*, New Delhi, India, pp. 1003–1007, 2019.

[17] W. Wu, K. Cai, Y. Yan and Y. Li, "An improved SVM model for flight delay prediction," in *2019 IEEE/AIAA 38th Digital Avionics Systems Conf.*, San Diego, CA, USA, pp. 1–6, 2019.

[18] M. Güvercin, N. Ferhatosmanoglu and B. Gedik, "Forecasting flight delays using clustered models based on airport networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 1, pp. 1–11, 2020.

[19] R. Yao, W. Jiandong and X. Tao, "Prediction model and algorithm of flight delay propagation based on integrated consideration of critical flight resources," in *2009 ISECS Int. Colloquium on Computing, Communication, Control, and Management*, Sanya, China, pp. 98–102, 2009.

[20] X. Dou, "Flight arrival delay prediction and analysis using ensemble learning," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conf.*, Chongqing, China, pp. 836–840, 2020.