Flight Arrival Delay Prediction And Analysis Using Ensemble Learning

Xiaotong Dou[1]
1. College of Statistics and Mathematics, Zhejiang Gongshang University
Hangzhou, China
douxiaotong0803@163.com

***Abstract*—With the development of the civil aviation transportation industry in recent years, the volume of civil aviation transportation has increased rapidly.Increased carrier costs and reduced airport operating efficiency caused by flight delays have become issues that need to be addressed. How to improve the accuracy of predicting flight arrival delay time is of great significance for improving airport transportation efficiency, rationally scheduling flights and improving passenger comfort. In this paper, the Cat-boost model is utilized on the U.S Domestic airline on-time performance data from U.S.Transportation Administration, combined with the characteristics of the model to determine the influencing factors, and to predict the arrival delays of flights within the United States.The accuracy;precision and some other criterion of the model are given to evaluate the performance on the data. A better effect is obtained: the accuracy reach 80.44% in this case. Finally, the specific delay time is predicted, we found that the support vector machine has the best prediction result for the flight delay time, the average prediction error is 9.733 min,which has a certain reference value for flight operation and airport scheduling.**

Keywords—Ensemble Learning ; Cat-boost ; Flight Delay ; Prediction

I. INTRODUCTION

In the process of civil aviation transportation, flight flight plans are affected by many factors such as weather, safety, and airport scheduling. Frequently, the flight schedule cannot be arrived at the scheduled time, which affects subsequent flight scheduling. Predicting flight delays in advance is of great significance for coordinating airport operations and improving airport efficiency.

Scholars at home and abroad have established prediction models from multiple perspectives based on the various factors affecting the uncertainty of flight operations:For example, Wen-BoDu [1] established a cause-and-effect network (DCN) of delay by using the cause-and-effect test of the Granger and analyzed the actual factors that affect flight delays. Alice Sternberg[2] use association rules to study hidden patterns of those delayed flights.In recent years, using machine learning and neural networks[3] to predict flight delay has become a major trend. BinYu[4] established a deep belief model to analyze and predict the internal model of flight delays. Poornima[5] used a non-parametric reinforcement learning (RL) based method to study the deviation between the planned taxi time and the actual time of the aircraft. Rodríguez-Sanz et al.[6], Cheng-Lung Wu et al.[7] predicts delays based on Bayesian network models, and combines prediction models with reliability models. RobertoHenriques[8] applied the multilayer perceptron on flight delay prediction to obtain a high accuracy rate. Sun Choi et al.[9] sampled the unbalanced flight data set and built a model based on Ada-boost, and got a relatively good results. A BN model that combining genetic algorithm(GA) with simulated annealing algorithm(SAA) was developed by WeidongCao et al. to predict the departure delay of a flight[10].

Traditional machine learning models have also achieved certain results in flight delay prediction. Methods such as decision tree, KNN, random forest [11], support vector machine and other methods have relatively high prediction accuracy on larger flight data.However, when using these methods to predict flight delays, a large number of features are often required for pre-processing. For example, the feature of flight takeoffs and landings requires one-hot encoding processing, which may cause the data to be too sparse. Encoding for airlines, aircraft tail coding, etc.is also required. A lot of information is lost during this process.With the rapid development of the aviation industry, the amount of flight data generated has increased dramatically, and traditional models are unable to satisfy the processing of large data sets. Due to the imbalance of data in the binary classification prediction, the model is prone to over fit in the application of actual data.

The Boosting model is an ensemble learning method used to improve the accuracy of weak classification algorithms. This method constructs a series of prediction functions and then combines them into a prediction function in a certain way. It continuously reduces the loss through a series of iterations of the model. Function values and improve the accuracy of the base classifier. Cat-boost is an open source gradient enhancement algorithm developed by Yandex Company (Dorogush et al.,2018). It allows users to quickly process the classification characteristics of large data sets, which can be used to solve regression; classification and ranking problems. Especially in the prediction of unbalanced data sets, which is superior to the performance of previous Light-GBM and XG-boost model.

Cat-boost uses a symmetric tree method, which makes up for the lack of previous boosting algorithms in processing category features, improves the robustness of prediction results, and has a high accuracy rate for the prediction of unbalanced data sets. In this paper, cat-boost is applied to the annual domestic punctual data set in the United States. Based on the grouped sampling data, the binary prediction is performed on whether the flight is delayed by more than fifteen minutes of CRS plan time.The accuracy rate of cat-boost is 80.44% after

978-1-7281-4390-3/20/$31.00 ©2020 IEEE

adjustment.Also we use cat-boost; Multilayer perceptron; bagging regression and support vector machine model to predict the specific delay time and compare the prediction error.

II. CLASSIFICATION PREDICTION BASED ON CAT-BOOST

A. CAT-BOOST MODEL

Gradient boosting is a useful machine learning technique that can achieve superior results in practical tasks like weather forecast and some binary problems with huge amount of features.It also have a great performance in the filed dealing with heterogeneous features, noisy data and complex dependencies.

Cat-boost is a novel gradient boosting technology proposed by Yandex Company (Dorogush et al., 2018). It's a machine learning framework based on gradient boosted decision trees,also combined with Logistic Regression Model.In each iteration, ordered boosting was used to modify the original algorithm's method of calculating loss functions and gradients based on the same data set. An unbiased estimate of the gradient was obtained, which effectively improved the generalization ability of the model.

The key steps of the model are as follows:

Whether the flight delays studied in this article is a binary problem. At this time, the loss_function of model is:

$$L(y, f(x)) = \log(1 + \exp(-yf(x)) \tag{1}$$

The value of the above loss function is in the range of (-1,1). It can be seen that the reverse derivative is:

$$r_{ti} = y_i / (1 + \exp(y_i f(x_i))) \tag{2}$$

An approximate estimate of each tree node is:

$$c_{tj} = \sum_{x_i \in R_{ij}} r_{ti} / \sum_{x_i \in R_{ij}} r_{ti}(1 - r_{ti}) \tag{3}$$

For categorical variables, cat-boost uses the ordered boosting method. That is, suppose the original categorical variable order is: $t = (t_1, t_2, ...t_n)$,Randomly traverse the whole random sequence, and calculate the value of the first p categorical variable records. $t_{p,k}$ can be present as :

$$\frac{\sum_{j=1}^{p} [x_{j,k} = x_{i,k}] \cdot Y_i + a \cdot P}{\sum_{j=1}^{n} [x_{j,k} = x_{i,k}] + a} \tag{4}$$

Among above, the prior value P and $a > 0$ is parameters to reducing noise in data set. In this way, categorical variables that do not have an order or level relationship can be well transformed.

B. IMPROVEMENT

Cat-boost's improvement of traditional encoding modes is mainly reflected in：

1.It is not necessary to encode categorical variables during the data pre-processing phase. It is processed automatically by the model during the training process. At present, target statistics is a processing direction that can minimize the amount of information loss in the encoding process.

2.Feature combinations were performed on all features. When the decision tree was split, a greedy algorithm was used to construct new and as many features as possible. All classification nodes in the tree were considered as new classification features, making full use of the data.

3.A symmetric tree is used as the base model. The traditional GBDT model uses target statistics to transform categorical variables into numerical variables, which will cause deviations between the transformed data distribution and the original data. To overcome this gradient bias, cat-boost proposes an ordered boosting method, which reduces the over fitting of model.In combination, the original data distribution is retained.

III. FLIGHTS DELAY PREDICTION

A. DATA PROCESSING

This paper choose all US domestic flight operation records provided by the United States Department of Transportation for the twelve months from October 2018 to September 2019. A flight delayed by 15 minutes or more are considered as delayed flights, and a flight delayed within 15 minutes or on time and early is considered a punctual flight. The features provided include plans; actual landing time information; landing city; airport information; aircraft information and so on.

First, simply processed was operated on this data set, records with too much missing values are eliminated, and variables such as actual arrival time, etc. obtained after the flight operation are removed. Due to the large amount of data, the data was grouped and sampled by month, and a total of 2153768 experimental data were obtained.

B. FEATURES SELECTION

The choice of features plays a key role in the final effect of the model. Different models need to select suitable features according to the characteristics of the model to achieve better results. The embedded feature selection method uses machine learning models for feature selection.The feature selection process is integrated with the learner training process. It can automatically make reasonable selections of features during the training process and give each feature a score.

Because cat-boost itself has the effect of scoring features, it is possible to select parameters that are more important to the model when the threshold is unknown, so here we use cat-boost to evaluate the features of each feature to select features,

and finally get the following 15 features to build a training model . Some features which have strong correlation with each other have been removed before the evaluation.such as actual elapse time for an aircraft ,and some time features that obviously imply to the delay of an airline.

Considering that the flight delay is subject to the actual delay of arrival, the two categorical variables ARR_DEL15 are used as prediction objects, where 1 represents delay and 0 represents on-time performance. The specific type and descriptions of selected influence factors are as follows:

TABLE I. SYMBOLS DESCRIPTION

variables	type	description
QUARTER	int64	Quarter (1-4)
MONTH	int64	Month
DAY_OF_MONTH	int64	Day off month
DAY_OF_WEEK	int64	Day of week
OP_UNIQUE_CARRIER	object	Code assigned by IATA and commonly used to identify a carrier.
TAIL_NUM	object	tail number
ORIGIN_AIRPORT_ID	int64	An identification number assigned by US DOT to identify a unique origin airport.
ORIGIN_CITY_NAME	object	City Name of origin airport
DEST_AIRPORT_ID	int64	An identification number assigned by US DOT to identify a unique destination airport.
DEST_CITY_NAME	object	City name of destination airport
CRS_DEP_TIME	float64	CRS departure time
CRS_ARR_TIME	float64	CRS arrival time
AIR_TIME	float64	Flight time, in minutes
DISTANCE	int64	Distance between airports (miles)

The feature importance given by cat-boost score can be counted as shown in Figure 2.

As can be seen from the above figure, the time characteristics have a large impact on the mode performance; the air time and flight distance will also have a greater impact on on-time performance of specific flight;Different carriers and specific aircraft will also have a slight influence of on-time performance.

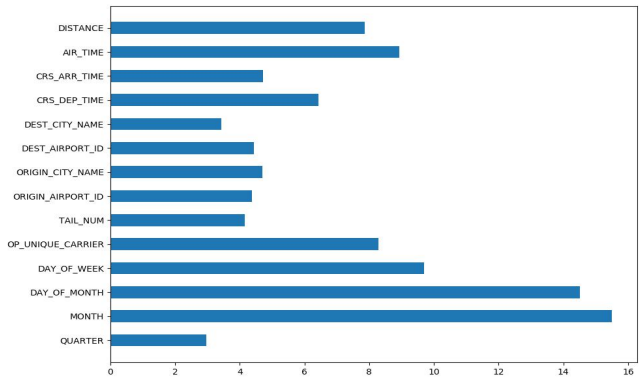

Fig. 1. Feature importance score for flight delay data

C. *FEATURE PRE-PROCESSING*

In the data of flight delay, variables such as carrier information, flight date, and aircraft tail number are all categorical variables and cannot be processed directly. Other variables are floating-point data and integer data that can be directly calculated. When using traditional machine learning models for prediction, you must first one-hot encoded these categorical variables. When using the catboost model for prediction, Ordered boost is introduced to encode the categorical variables, and only the categorical variables need to be labeled when the data is introduced into the model.

D. *EVALUATION CRITERIA*

The evaluation criteria of the model are very important for the measurement of the final result. Different scenarios often have different focuses. In general, the pros and cons of a classifier can be evaluated using accuracy, precision, and recall. Accuracy refers to the proportion of all samples that are correctly classified. Accuracy refers to the proportion of samples that are judged to be positive. The recall ratio refers to the proportion of samples that are actually positive. Because the probability of flight delay is far less than the probability of on-time arrival of the flight, in actual research, we tend to pay attention to the accuracy of on-time flights, that is, the precision and accuracy. At the same time, the accuracy of the model must be higher than the accuracy obtained when all the samples are judged as the majority of samples, so that the model efficiency is better than the random judgment result,witch called ZeroR Classifier. ZeroR Classifier set an reference for classification to make sure that the prediction accuracy is higher than classified all the variables to be explained into the category with the largest proportion.

TABLE II. CONFUSION MATRIX

	Positive	**Negative**
True	TP	TN
False	FP	FN

In the binary classification problem, the selection of the decision threshold has a huge influence on the quality of the

classifier. It is a more common method to determine the value of the decision threshold through the ROC-AUC image and FPR-FNR image. The ROC curve refers to the receiver operating characteristic, and each point on it reflects the susceptibility to the same signal stimulus. The horizontal axis is the negative positive class rate FPR, which represents the probability of dividing a true negative sample into positive samples, the vertical axis is the accuracy rate, and the AUC is the area formed by the ROC and the x axis. The FPR-FNR curve shows the probability of incorrect classification of positive or negative samples when taking different thresholds.

According to the requirements of the binary forecast of Airline On-Time Statistics data and the unbalanced data characteristics of the samples, AUC-ROC and F1 were selected as the model's pros and cons. At the same time, the ZeroR is used as the threshold to check the validity of the model. The evaluation criteria involved above are calculated as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FP} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$FNR = \frac{FN}{TP + FN} \tag{9}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{10}$$

$$ZeroR = CorrentlyClassifiedIns\tan ces = 68.65\% \tag{11}$$

Among them F_1 is the weighted harmonic average of precision and recall, the value range is $F_1 \in [0,1]$, and when it is 1, the model is optimal. This experiment is based on Python 3.6 version ,sklearn and Cat-Boost package for practice.

The original test data includes 519,734 positive records for punctual flights records and 126,397 negative records for delayed flights.Use 70% of the data as training data and 30% as test data. The results of the confusion matrix predicted by the cat-boost model for test data are shown in the following table:

TABLE III. CONFUSION MATRIX FOR DELAY PREDICTION

	Positive	**Negative**
True	511792	7942
False	106415	19982

The AUC-ROC and FPR-FNR images of the model on the test data set after adjusting the parameters are shown below:

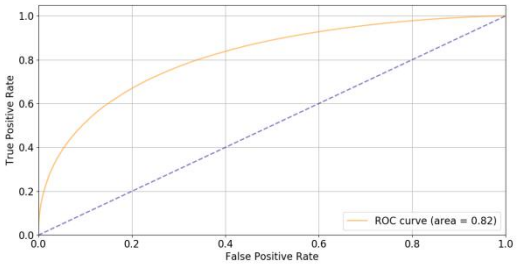

Fig. 2. AUC-ROC curve of flight delay prediction

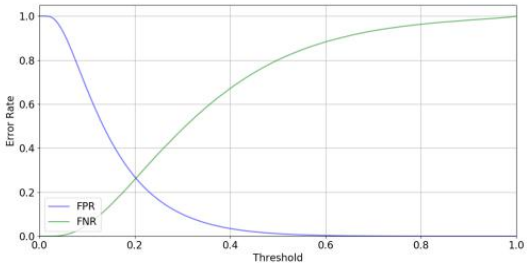

Fig. 3. FPR-FNR curve of flight delay prediction

In order to compare the prediction results obtained by different threshold values, the threshold values are calculated at intervals of 0.1 between 0 to 0.3 to calculate accuracy and F1. The results are shown in the following table:

TABLE IV. ACCURACY AND F1 AT DIFFERENT THERESHOLD

threshold	**0.1**	**0.2**	**0.3**
Accuracy	0.4544	0.8044	0.6325
F1	0.2833	0.8995	0.5812

Comprehensively evaluate the effect and select 0.2 as the prediction threshold.The accuracy is 80.44%,which is higher than the standard of Zero-R,which means the model have better performance than random choose. The prediction standard values under the optimal prediction threshold are shown in the following table:

TABLE V. .EVALUATION PARAMETERS FOR PREDICTION

Evaluation parameters	**Value(%)**
Accuracy	80.44
Precision	98.47
Recall	82.79
FPR	20.01
FNR	3.76
F1	89.95

IV. Flight Delay Duration Prediction

For flight delays, just to know whether the delay is delayed is not enough. Predicting the specific estimated delay time in advance is also of great significance for improving airport dispatching efficiency and controlling flight operating costs. In order to make an effective prediction of the specific time delay of the flight, we use several common regression prediction algorithms to predict the delay at the same time for the round-trip flight between John F. Kennedy International Airport and O'Hare International Airport. Arrival delay time is selected as the dependent variable,the delay time are presented in minutes, which is expressed as a negative number for the flights that arriving early; 12 influencing factors such as quarter,carrier, and distance are independent variables, with a total of 5,375 valid records in this data.The specific flight delays time pattern are shown below:

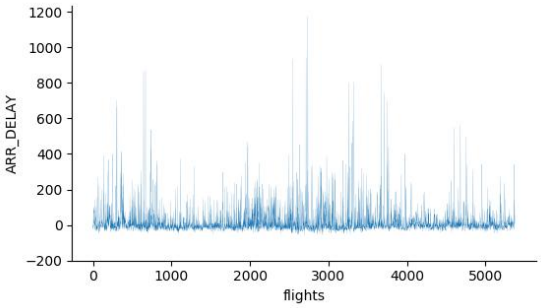

Fig. 4. Delay Performance between JFK and ORD

Considering the departure and arrival airports have fixed parameters and a small amount of data, one-hot encoding processing is performed on classification variables such as carriers, tail number, departure and arrival cities. Use grid search and cross-checking methods to determine the optimal parameters and at the same time fit the data in selected models. With MAE and RMSE as the standard for measuring the pros and cons of the model, the fitting effect of each method on the test set is as follows:

TABLE VI. Errors of delay prediction

	MAE	**RMSE**
Cat-boost	16.348	13.495
MLP Regression	21.956	23.126
Bagging Regression	14.241	12.229
SVR	9.733	8.245

As it can be seen from the above table, with the control of independent variables and data consistent, the support vector machine(SVR) model achieved lowest mean average error(MAE) and root mean square error(RMSE) in flight delay prediction. The time prediction of delay flights have a certain reference value for airline scheduling and airports arrangement.

V. Conclusions

The prediction of flight delays is of great significance for improving the operating efficiency of airports and reducing the operating costs of airlines. According to the results of the above indicators, it can be seen that cat-boost can better deal with categorical variable encoding problems and imbalanced data set over-fitting problems in flight delay data, and use the embedding method to sort and analyze the factors affecting flight delays. After adjusting parameters, the accuracy and precision of the flight delay prediction obtained are better than that of traditional machine learning models, and it has better robustness. In the data processing stage, the sorted data sorting method of ordered boost avoids the coding problem of categorical variables and reduces the information loss rate of categorical data. The support vector machine was proved have a better performance on specific time prediction of delay flights,which have a mean error of 9.733 minutes.However, due to the lack of data and other issues, detailed weather and aircraft data cannot be collected ,the accuracy of the model needs to be further improved.

REFERENCES

[1] Du, Wenbo & Zhang, Ming-Yuan & Zhang, Yu & Cao, Xian-Bin & Zhang, Jun. (2018). Delay causality network in air transport systems. Transportation Research Part E: Logistics and Transportation Review.

[2] Alice Sternberg, Diego Carvalho, Leonardo Murta, Jorge Soares, Eduardo Ogasawara,2016,An analysis of Brazilian flight delays based on frequent patterns,Transportation Research Part E: Logistics and Transportation Review.

[3] Sina Khanmohammadi, Salih Tutun, Yunus Kucuk,2016,A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport,Procedia Computer Science.

[4] Yu, B. , Guo, Z. , Asian, S. , Wang, H. , & Chen, G. (2019). Flight delay prediction for commercial air transport: a deep learning approach. Transportation Research Part E: Logistics and Transportation Review.

[5] Poornima Balakrishna, Rajesh Ganesan, Lance Sherry,2010,Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures,Transportation Research Part C: Emerging Technologies..

[6] Álvaro Rodríguez-Sanz, Fernando Gómez Comendador, Rosa Arnaldo Valdés, Javier Pérez-Castán, Rocío Barragán Montes, Sergio Cámara Serrano,2019,Assessment of airport arrival congestion and delay: Prediction and reliability,Transportation Research Part C: Emerging Technologies.

[7] Cheng-Lung Wu, Kristie Law,2019,Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model,Transportation Research Part E: Logistics and Transportation,Review.

[8] Roberto Henriques, Inês Feiteira,2018,Predictive Modelling: Flight Delays and Associated Factors,Hartsfield-Jackson Atlanta International Airport,Procedia Computer Science.

[9] Choi S , Kim Y J , Briceno S , et al. Prediction of weather-induced airline delays based on machine learning algorithms[C]// Digital Avionics Systems Conference. IEEE, 2016.

[10] [10]Weidong Cao, Xiangnong Fang,2012,Airport Flight Departure Delay Model on Improved BN Structure Learning,Physics Procedia.

[11] Yu B , Wang H , Shan W , et al.2017, Prediction of Bus Travel Time Using Random Forests Based on Near Neighbors[J]. Computer-Aided Civil and Infrastructure Engineering.