Flight Delay Prediction using Temporal and Geographical Information

Jianmo Ni
UC San Diego
jin018@ucsd.edu

Xinyuan Wang
UC San Diego
xiw193@ucsd.edu

Ziliang Li
UC San Diego
zil264@ucsd.edu

ABSTRACT

Flight delay is one of the important measurements of airline services. In this paper, we study a public dataset to analyze the flight delay in the United States. Temporal and geographical data are examined to determine which are most predictive of flight delays. Six classification models are implemented to classify a flight into potential delayed one or non-delayed one. Weighted loss function and ensemble models are introduced to improve the performance on the imbalanced dataset. Ripple effect is also discussed and temporal delay feature is extracted to further improve the models.

KEYWORDS

flight delay prediction, imbalanced dataset, weighted loss function, temporal and geographical data

1 INTRODUCTION

With the advance of global economy, traveling by air have become a natural choice for traveling purpose or business purpose. While enjoying the high efficiency and comfortable service accompanied by taking aircraft, we are suffering from its typical shortcoming, potential and unexpected unpunctuality. Possible unpunctuality seems to be unique to air travel as we do not always see it on the other two most popular choices for long-distance travel. Despite people can't prevent a delay, we are always eager to know whether a flight will be delayed even just for emotional comfort. In this paper we work on the prediction for whether a flight will be delayed or not.

In this project, we use the dataset 2015 Flight delays and cancellations from Kaggle, which contains all flight information and airport information in 2015 of USA. To simplify, we only use part of the dataset. In Sec. 2 we discuss the related work in this area. In Sec. 3 we analyze the dataset to discover its essential properties so that we can have a better choice for features used in the prediction model. The prediction task is described in Sec. 4 and the metrics to evaluate model and the baseline model in Sec. 5. We also introduce the features used and how to get them. In total, we use six models, including Naive Bayes, Logistic Regression, Weighted Logistic Regression, Random Forest, Weighted random forest and the ensemble of Random Forest and Logistic Regression. We compare these models and analyze the strengths and weakness. Finally, we make a conclusion on the project in Sec. 7

2 LITERATURE

Flight delay is increasing and leads to extra costs to US air travel systems, airlines and passengers. The total delay impact study [1] analyzes a broader consideration of possible cost components caused by flight delays. Since it becomes a serious problem in US, analysis and prediction of flight delay are being studied in order to reduce the large costs.

Mueller [3] analyzed the departure and arrival data for ten major airports in US and characterize the delay data in a statistical method. It revealed that the ground movement inefficiency contributed most to surface delays and weather was the main cause of delay for air traffic control. The departure and arrival delays were modeled with probabilistic demand forecasting methods. Statistical approaches were also applied in [5] with long-term trend and short-term pattern. The key component of this model is the estimation of the delay propagation effect, where delay propagation was known as delay built up from previous flights and effects on delays. Rebollo and Balakrishnan [4] proposed a new class of models using machine learning techniques with air traffic network characteristics to predict air traffic delays. The proposed models considered both temporal and spatial delay states as explanatory variables, and used Random Forest algorithms to predict departure delays 2-24 h in the future.

Deep learning has also become a promising method for solving data analytics problems such as traffic flow prediction. Kim and Choi (2016) [2] selected the recurrent neural networks for the day-to-day delay status prediction, which first trained the model for a reliable delay status of a single day then the delay states of individual flights of that day. Deep learning algorithms capture the sequential and temporal relationships in the day-to-day delay data

3 DATASET ANALYSIS

We get the dataset from Kaggle. It comprises of information about flights in 2015 of USA and airports. Their sizes are over 5 million and 322 respectively. To simplify the analysis and facilitate model development, we randomly extract 1 million data about flights. The flights information consists of 31 fields, basically about the date and the time of departure time and arrival time, the origin and destination airports, flight number, distance, taxi-in time, taxi-out time, elapsed time in air, delay time, whether cancelled, distribution of delay time among multiple causes. The airports information contains 7 fields, including IATA code, name, city, state, country, latitude, longitude. Airlines information consists of IATA code and the full name of the airline.

The Figure 1 and Fig. 2 indicate the distribution of all airports across the United States as well as their delay rate (the ratio between number of delayed flights to number of total flights) and cancel rate (the ratio between number of cancelled flights to total flights). Each spot indicates an airport and its color indicates the value of delay rate or cancellation rate. The color gradient box in Fig. 1 and 2 represents the scope of real values. As can be seen, an airport that has a greater delay rate also has a greater cancel rate in most cases, which means some airports tend to delay flights for

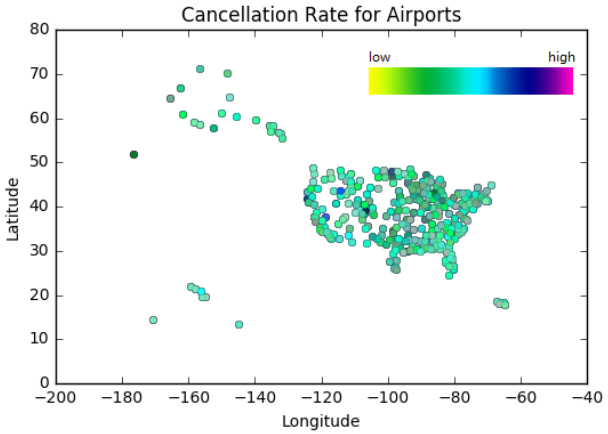

Figure 1: Flights cancel rate distribution of all the airports across US.

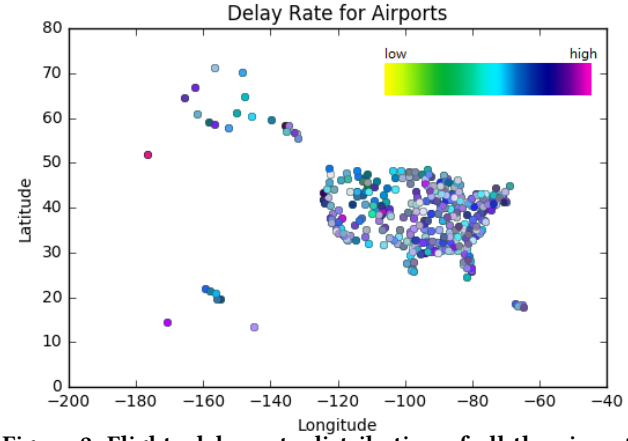

Figure 2: Flights delay rate distribution of all the airports across US.

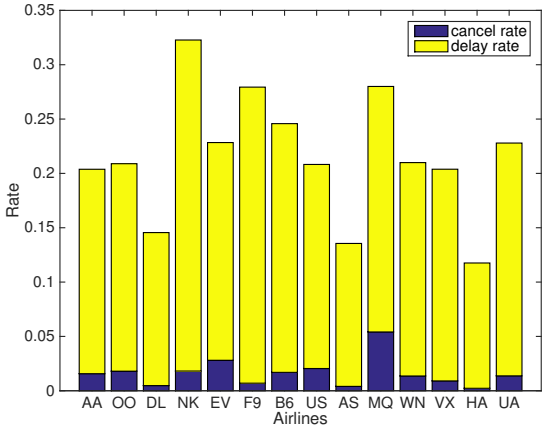

Figure 3: Flights cancel and delay rate distribution of all the airlines.

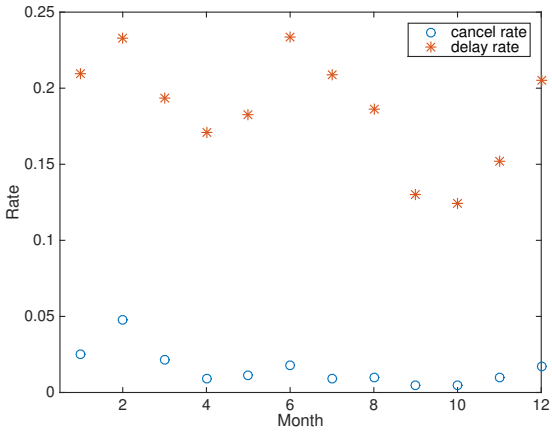

Figure 4: Flights cancel and delay rate distribution versus month.

the possible reasons that they are less efficient in security check or runway clearance. Therefore, whether a flight will be delayed may be associated with which airport it departs from.

Fig. 3 describes the cancel rate and delay rate of each airline. Customers are suggested to choose airlines with a rather low delay/cancel rate so that the flight is more likely to arrive on time or even in advance. So we consider the airline to which a flight belongs as an important feature.

We also consider month as a feature. Fig 4 illustrates the trend of delay/cancel rate through months. Delays and cancellations happen more frequently in some months from December to February and from June to July. It is probably because the temperature is relatively higher or lower in this time. A more interesting trend about departure time is presented in Fig. 5. An strongly positive linear relation happens between 5 am to 8 pm. Then we see a decline from 8 pm. Hence departure time is considered as a potentially influential factor. The day of week in Fig. 6 also shows a variance. A great difference is seen on Monday and Saturday. Perhaps it is because they are the first day to work and the first day to rest.

4 PREDICTIVE TASK

4.1 Task Description

Our task is to determine whether a flight is *delayed* or *non-delayed*. We label flights with ARRIVAL_DELAY greater than 15 minutes as *delayed* flights and flights with ARRIVAL_DELAY smaller than 15 minutes as *non-delayed* flights. The loss function is defined as follows. Here the weight term w_1 and w_0 are introduced to balance the dataset for positive samples and negative samples, respectively.

$$L = \frac{1}{N}\sum_{i=1}^{N} w_1 y_i log(p_i) + w_0(1-y_i)log(1-p_i) \quad (1)$$

4.2 Data Preprocessing

We explore correlation between different features and labels from our exploratory analysis. The following features show correlation with the arrival delay and hence are considered in the classification models.

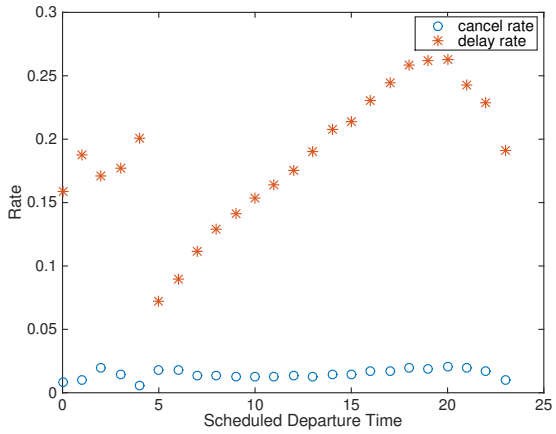

Figure 5: Flights cancel and delay rate distribution of scheduled departure time.

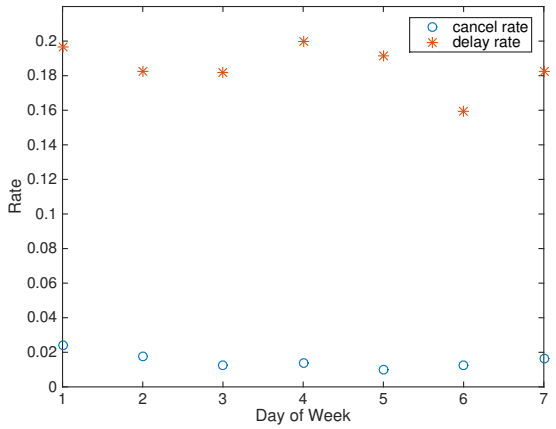

Figure 6: Flights cancel and delay rate distribution versus day of week.

To process the data, features with discrete values are first converted into categorical variables using one-hot encoding.

Moreover, we manually generated a new feature ARRIVAL_DELAY-_AVG by calculating the average arrival delay rate for each flight on the training set.

Finally, there are 734 features and we store them in sparse matrix to reduce the memory usage.

Discrete:

- MONTH
- DAY
- DAY_OF_WEEK
- AIRLINE
- ORIGIN_AIRPORT
- DESTINATION_AIRPORT
- SCHEDULED_DEPARTURE

Continuous:

- DISTANCE
- AVG_ARRIVAL_DELAY

Table 1: Importance features of Random Forest

Feature	Importance
ARRIVAL_DELAY_AVG	0.110780
DISTANCE	0.047330
SCHEDULED_DEPARTURE_6	0.034319
SCHEDULED_DEPARTURE_7	0.023623
AIRLINE_DL	0.019887

Since the dataset is large, we apply conventional validation method rather than cross-validation. We divide the whole dataset into train, validation and test set by 0.8,0.1 and 0.1 proportion.

5 MODELS

5.1 Baseline

For the binary classification task, we first implemented Naive Bayes Model as Baseline. As shown in table 2, it has an accuracy of 0.403 and reaches F1 score 0.46 and 0.33 on non-delay and delay flights.

In the following experiments, we would consider both accuracy and F1 score on delayed flights as evaluation metrics. As discussed later, the data is imbalanced and we would pay more attention to the prediction of delayed flights.

5.2 Predicting with Single Classifers

We then studied on Logistic Regression (LR) and Random Forest (RF), respectively. We first trained the two models with conventional loss function (rather than the weighted loss function defined in this paper).

As table 2 shows, both classifiers have accuracy near 0.8 and precision on non-delayed flights near 1. Whereas, they have near 0 recall and F1 score on delayed flights, which means they classify only a small number flights as delayed.

The reason is that the imbalance of the dataset makes the predictor to classify all flights as non-delayed so as to decreased loss function. According to the analysis before, the ratio of the positive samples (delayed flights) and the negative samples (non-delayed flights) are around 0.25.

To solve the imbalance problem, we apply weighted loss function as defined in equation 1. Based on the observed ratio of positive and negative samples, we assign $w_1 = 4$ and $w_0 = 1$ in our experiment.

$$\frac{w_1}{w_0} \approx \frac{\#negative\ samples}{\#positive\ samples} = 4$$

As table 2 shows, the F1 on delayed flights increase to 0.32 and 0.31 on LR and RF respectively. On the other hand, the weighted term reduces the accuracy compared to the non-weighted models. However, the accuracy (0.625) is 55% better than the baseline (0.403) while the F1 scores are almost the same.

5.3 Predicting with Ensemble Models

As discussed in the previous sections, individual models do not achieve accuracy more than 0.625 on the imbalanced dataset while maintaining F1 score. In this subsection, we try to implement an ensemble model to improve the accuracy. We focus on a cascaded RF and LR model.

Table 2: Performance of classifiers for non-delayed and delayed flights

Model	Non-delay flights				Delayed flights			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Naive Bayes (Baseline)	0.403	0.84	0.32	0.46	**0.403**	0.21	0.74	**0.33**
Logistic Regression	0.802	0.81	1.00	0.89	0.802	0.25	0.01	0.01
Logistic Regression (weighted)	0.566	0.83	0.57	0.68	0.566	0.23	0.53	0.32
Random Forest	0.804	0.80	1.00	0.89	0.804	0.31	0.00	0.00
Random Forest (weighted)	0.625	0.83	0.67	0.74	**0.625**	0.24	0.43	**0.31**
Ensemble Model	0.547	0.83	0.55	0.66	0.547	0.23	0.55	0.32

Table 3: Performance of classifiers for non-delayed and delayed flights at SFO airport

Model	Non-delay flights				Delayed flights			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.806	0.84	0.94	0.89	0.806	0.52	0.26	0.35
Logistic Regression (weighted)	0.672	0.89	0.68	0.77	**0.672**	0.33	0.66	**0.44**
Random Forest	0.804	0.82	0.98	0.89	0.804	0.52	0.1	0.17
Random Forest (weighted)	0.763	0.85	0.85	0.85	0.763	0.40	0.40	0.40

We first use a weighted RF model to perform further feature selection based on the importance of each original features. Then we only maintain the selected features and train on the weighted LR model.

After training on the weighted RF model, the most important 5 features are selected as table 1. According to the feature importance, the top three features are ARRIVAL_DELAY_AVG, DISTANCEand SCHEDULED_DEPARTURE_6.

Intuitively, the average arrival delay gives information about the history delay and works as the most crucial factor. DISTANCE reflects how long it will take to reach the destination. The longer the distance is, the higher possibility that some urgency would happen to cause the delay. According the explanatory analysis, morning flights are less likely to be delayed and hence it is also useful to predict delay.

Subsequently, we generate new train data and put it into LR model. The new train data is significantly reduced in size and can be trained in a short time. Performance is shown in table 2. The accuracy is decreased while the F1 score remains the same as weighted LR model. The performance is not improved which shows the problem with the feature engineering. The selected by RF cannot provide enough predictive power.

6 EXTRA EXPERIMENT

To further explore the dataset, we decided to conduct an extra experiment on a subclass of the dataset. We focused on the San Francisco Airport (SFO) and built models to predict flight delay of that airport.

Besides the features mentioned in the previous section, we tried to look at the average arrival delay of the previous day for each flight. Intuitively, flight delay is greatly influenced by the 'ripple' effect and we expect the delay situation of yesterday would have great impact on the flight delay.

As shown in table 3, all models show significantly improved accuracy and F1 scores for delayed flights compared with the total dataset. The weighted LR model reaches the best F1 score of 0.44 and with an accuracy of 0.672. Though the extra experiment is conducted on the subset of flights at SFO airport, the improvement in performance can also show the importance of the feature "Average Arrival Delay of yesterday".

7 CONCLUSIONS

This paper studied the flight delay prediction problem. In the paper, we introduced weighted loss function to deal with the imbalance of dataset. We implemented Naive Bayes, Logistic Regression and Random Forest to classifer delayed fligts. The experiment results show the importance of weighted loss function and feature engineering. The flight delay is time-dependent and how to efficiently use the short-term temporal data will greatly impact the performance.

ACKNOWLEDGMENTS

The authors would like to thank Professor Julian McAuley for providing the supervise throughout this quarter.

REFERENCES

[1] Michael Ball, Cynthia Barnhart, Martin Dresner, Mark Hansen, Kevin Neels, Amedeo Odoni, Everett Peterson, Lance Sherry, Antonio A Trani, and Bo Zou. 2010. Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States. (2010).

[2] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. 2016. A deep learning approach to flight delay prediction. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*. IEEE, 1–6.

[3] Eric Mueller and Gano Chatterji. 2002. Analysis of aircraft arrival and departure delay characteristics. In *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*. 5866.

[4] Juan Jose Rebollo and Hamsa Balakrishnan. 2014. Characterization and prediction of air traffic delays. *Transportation research part C: Emerging technologies* 44 (2014), 231–241.

[5] Yufeng Tu, Michael O Ball, and Wolfgang S Jank. 2008. Estimating flight departure delay distributionsâĂŤA statistical approach with long-term trend and short-term pattern. *J. Amer. Statist. Assoc.* 103, 481 (2008), 112–125.