Predictive analysis of flight departure delays at the San Francisco national airport

Wildan Abdussalam
Home sweet home, Reitbahnstrasse 35, Dresden, Germany

We investigate factors that determine delays of flight departures at the San Francisco airport. The mean departure delays through 2016 as well as percentage of departure delays of carriages are provided to show seasonal traffic and airlines performances, respectively. Delay and no-delay departure flights are classified to construct the prediction of delay flight departure. We show that classification by means of Random Forest provides ≈ 81% accuracy prediction.

INTRODUCTION

Predictive analysis based on supervised learning has been not only developed in recent years but also applied to broad range of fields of both natural and social sciences. For cases of the latter field, this can serve as a tool to predict fascinating economic, political and social problems. An example of them was flight delays that might costed tremendous loss. In this paper, the analysis of departure flight delays at the San Francisco national airport is provided.

INPUT AND FEATURE SELECTIONS

To precisely provide the analysis, the data through 2016 were taken from United States Department of Transportation [1]. As the downloaded data were provided monthly, it requires a trick to import the data into PostgreSQL database with adjustable specifications. This can be efficiently performed with spark [3] and enable us to perform larger data tabulations from the old years such as 1980. After downloading, filtering, and processing them, the data were eventually saved in parquette files as there were several advantages of this technique: (i) the downloaded data could be easily extended to bigger size of data. For example, with this script one can download the data prior to 1980; (ii) since the data were saved to separated files, this enabled parallel-complex data processing which could be efficiently solved using Hadoop [4]. Should we need to show insitu big data visualisation, hdfs could be integrated easily with opensource projects such as OpenMd-api [5] as well as $Yt+$ [6]. After saving into parquette files, the data can be served as meaningful datasets. Next, pandas can be used to efficiently explore the informations within the data.

As the main objective of this assignment is to provide much information concerning the departure flight delays from San Francisco airport, we consider 10 features: *month, day of week, day of month, carrier, origin airport, destination airport, departure time, taxi out, distance, and delay for more than 15 minutes.* From those selected data there were 36445 rows of flight delays and 133485 no flight delays data, respectively. The chosen features were considered based on intrinsic discrepancies between two probability distributions (see Fig. 3 as an example):

$$\delta\{p_1, p_2\} = min\left\{\int p_1(x)log\frac{p_1}{p_2}dx, \int p_2(x)log\frac{p_2}{p_1}dx\right\}, \tag{1}$$

where the results are shown in the following table,

Departure time	0.3483067
Carrier	0.03462886
Destination	0.006120483
Month	0
DayofMonth	0
DayofWeek	0
Taxiout	0
Distance	0

This is aimed at not only gaining more informations of departure flight but also constructing a predictor of binary delay variable that can predict delay as well as no delay departure flights. In selecting methods, we compare the precisions of random forest algorithm to other methods such as logistic regression, decision tree classifier, and neural networks. The results will be shown later in Sec. .

METHODS

As mentioned in Sec. , the random forest will be applied to predict both the delay and no delay departure flights. In this section, we describe the technical method to subsequently explore the data. We first used the training set, after 70:30 split, with 10 features to train the supervised learning methods. Here we employ four methods: random forest, decision tree classifiers, logistic regressions and neural network. The implementation of those methods is available in scikit library [7]. The precision results of those methods are compared to choose the suitable method. As the random forest is selected, a grid search method is employed to improve its performance.

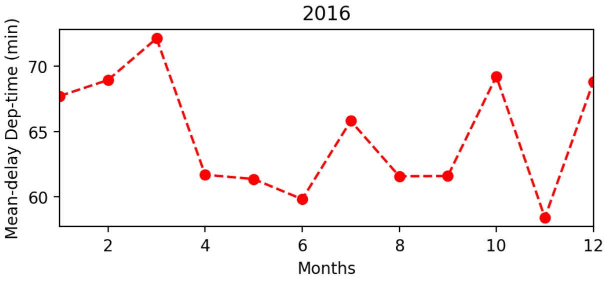

FIG. 1. (color online) The mean of departure delays through 2016.

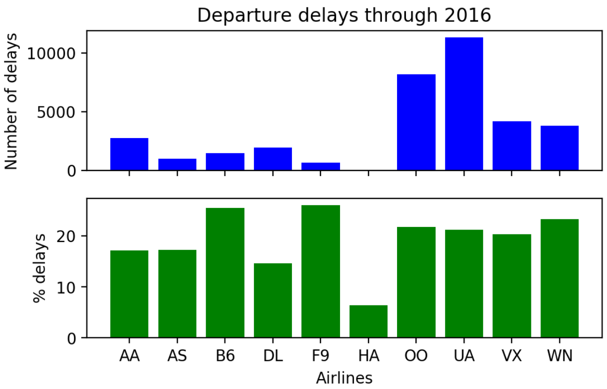

FIG. 2. (color online) Panel above shows the departure delays of airlines from San Francisco in 2016. The top figure shows the frequent number of delays against various types of carriers. The bottom one shows the percentage delays against various types of carriers. The percentage delays were defined as a ratio of 15 minutes departure delays to all departures of flights. The carrier codes can be found in the following link [2].

RESULTS AND DISCUSSION

The extracted data revealed that the San Francisco airport had been one of the busy airports in the United States. In 2016, there were 10 various airlines and 3236 flight numbers which departed to 81 destinations. Consequently, the airport sometimes faced traffic which led to the flight departure delays. Surprisingly, the delay could take up to 35 hours which might cause tremendous loss for either passengers or airlines. Fig. 1 shows the flight departure delays through 2016. In this case, only the flight delays more than 15 minutes were considered. It shows that high-mean delay of departure occurred in March, October and December. This delay can be related to the airline performances. Fig. 2 shows the airlines performances departed from the San Francisco airport through 2016.

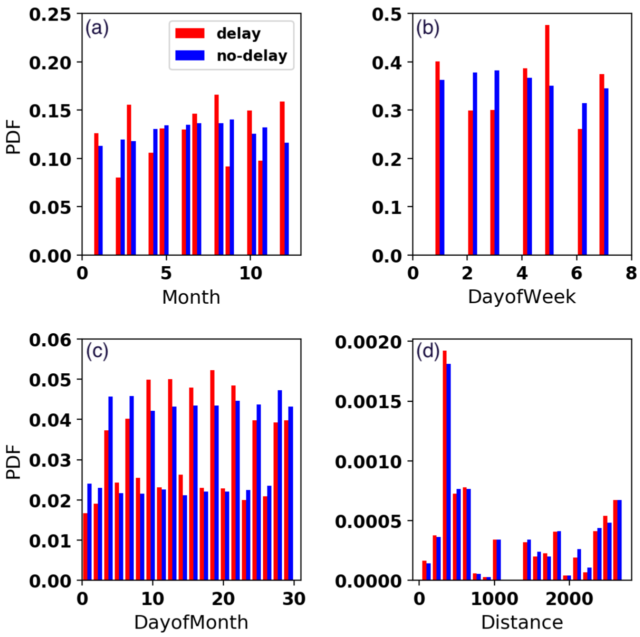

FIG. 3. (color online) Panels shows the probability density functions through (a) the months, (b) the days of week, (c) the days of month and (d) distances.

Fig. 2 top shows that the United airlines were the most departed-delay airlines through 2016 followed by Skywest and Virgin America airlines, respectively. This, however, unlikely to lead to the conclusion that the United airlines was the worst performance airlines through 2016. They conducted quite frequent flights in 2016 which label them as one of busiest airline in USA. Fig. 2 bottom demonstrates that in spite of the highest number of delays, the American airlines were not the worst performance airlines through 2016. Frontier airlines, however, placed a highest rank as the worst performance due to their less frequent flights. The best performance airlines were granted to Hawaiian airlines followed by Delta airforce as well as American airlines. This is due to the fact that these airline conducted less frequent flights than the aforementioned cases.

In addition to carrier, to gain some insights into more features, normalized distributions for the 'delay' and 'non-delay' cases as the function of various features were plotted. Fig 3 shows 4 examples of normalized distributions for the 'delay' and 'non-delay' cases as the function of four features. As shown in Fig 3 (a), the delays frequently occurred in March and December. Note that in contradiction to the Fig. 1 where the average of more than 15 minutes delays are considered, the distribution data provides the most frequent delay in August that relates to the end of summer vacations. Moreover, the departure delays frequently occurred on Friday [see Fig 3 (b) and (c)] that was associated with weekly-short dayoff. As the distances taken by the flight prone to departure delays were less than 1000 Km [see Fig 3 (d)], this

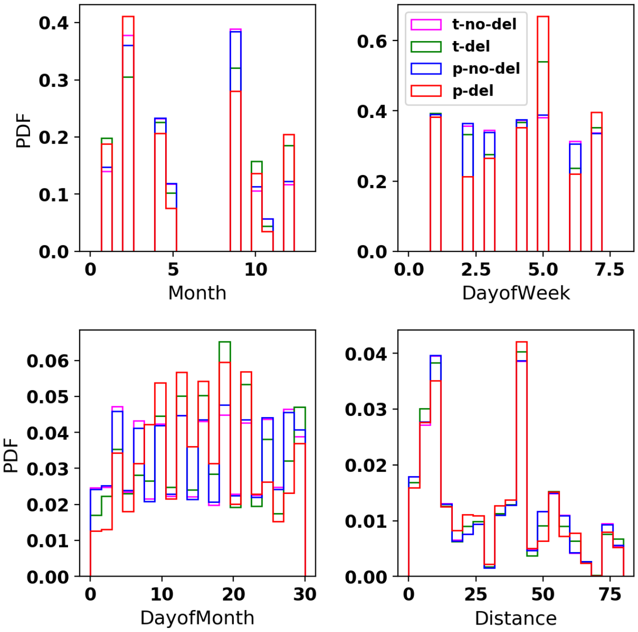

FIG. 4. (color online) Panels shows distribution trough the months, the days of week, the days of month and distances. T-no-del and t-del denotes data test for no delay and delay flights, respectively. P-no-del and p-del denotes prediction data for no delay and delay flights, respectively.

shows that most of passenger travelled from San Francisco to other destinations for visiting their relatives, returning back from work, or having a short-vacation break.

	random forest	decision tree classifier	logistic regression	neural networks
precision	0.818	0.752	0.814	0.815

As mentioned in Sec. , one can use these features to train the supervised learning methods. Therefore, we provide a comparison of precision results from the aforementioned methods (see table). An example of the random forest prediction results are shown in Fig. 4. Although the random forest provides the precision > 0.8, discrepancies remain apparent for each feature. Last but not least, the importance of feature which was taken from the predictive data test for delay and no delay are shown in the following table,

Departure time	0.1017827
Carrier	0.02940339
Destination	0.007990665

It shows that the departure time is the most essential factor that can affect the flight delays.

CONCLUSION

We have investigated the prediction of the flight departure delays at the San Francisco airport. Among three factors that determine the flight delays, the departure was an essential factor that can cause the flight delays.

[1] Bureau of transportation statistics. 2016. .
[2] Airline codes
[3] Spark
[4] Hadoop
[5] OpenMd-API
[6] Yt
[7] Scikit library