FLIGHT DELAY PREDICTION

- Predicting whether and how long one could expect a flight to delay

Final Project for

201903_Introduction to Data Mining_DATS_6103_12

Taught by professor Edwin Lo

By:
Olatunji Ayobami
Denny Fang
Sanat Lal
Vishal Pathak

Table of Contents

Predicting Airplane Delay 2
- Executive Summary 2
- Dataset Used 2
- Methodology Used 3
- Model Building Process 3
  - Data Cleaning 3
  - Initial Data Exploring 6
  - Model Building 7
- Conclusion 8

Table of Figures

Figure 1 Columns Information 5
Figure 2 Original Data Types 5
Figure 3 Columns with Null Values 6
Figure 4 Four Examples of Data Density Plots (Distribution) 7
Figure 5 Dependent and Independent Variables 8
Figure 6 Linear Model Accuracy 8
Figure 7 Confusion Matrix for logistic Regression 10

Predicting Airplane Delay

Executive Summary

Flight delays could always be annoying, especially in the case when the period of delay was so long that there was even a danger to miss the next flight. However, if there was a way to predict whether there would be a delay or even better – how long the delay could be, then people could make earlier preparation to reschedule following flights in an earlier manner. For that consideration, we adopted a dataset containing airline delayed time and other airline-related information provided by Kaggle to building a model, mainly aiming to solve the following questions: 1. Whether there would be a delay with certain publicly reachable resources; and 2. How long delayed time one could expect with the same information given. We deployed python sklearn and pandas library to build our model, and evaluate our model based on R-Square for linear regression and accuracy rate for logistic regression. As a brief result of our project, we found, it would be helpful to use the following factors in evaluating our model: week, month, airline carrier reference, planned elapsed time (in air time), distance between two departure and destinations, flight planned departure time, departure airport code, and taxi-in and taxi-out time.

Dataset Used

We chose the “Airlines Delay” data from www.kaggle.com/datasets[1], which was actually provided by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS), that tracks the on-

[1] Full dataset see: https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled and diverted flights appears in the data. We initially intend to use the whole dataset covering 10 years to build our model, however, due to our limited computation resources, we had to cut the dataset of the latest dataset of the year or 2018. In the dataset, 7213446 rows included, and 27 variables involved.

Methodology Used

Due to the huge data (7213446 lines of data included), it would both be impossible and impractical for us to manually explore and find patterns between flight delay information and related influencers. Therefore, we decide to first manually clean the data, and then adopt machine learning with Python sklearn library. For data cleaning, we firstly change data-types of certain columns with “object” type, and replace ‘null’ values with certain values, to make the data suitable for machine learning. Afterwards, used pandas, seaborn and matplotlib to make initial exploration in order to find some intuitive relationship between variables. Finally, we deploy machine learning method to dig out factors and their correlation with flight delays – to be specific, we used linear regression to predict the expected delay time of flights and used logistic regression to predict whether a flight might or might not be delayed.

Model Building Process

Data Cleaning

For Logistic Regression Model

No data mining projects could be finished without thoroughly understand the data first. So, in order to better understand data, we start our project by exploring the data first. We found the original dataset includes 28 attributes/columns (shown in figure 1), and while most of the data were in float format, some of them were object types (shown in figure 2). In addition, as shown in Figure 3, there were also many null values in the original datasets. So, we need to first clean the columns with null values and change data types of objects into suitable types (mostly integers) for the convenience of machine learning.

```
In [7]: len(airline.columns)
Out[7]: 28
```

Figure 1 Columns Information

```
In [20]: a = airline.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7213446 entries, 0 to 7213445
Data columns (total 28 columns):
FL_DATE                object
OP_CARRIER             object
OP_CARRIER_FL_NUM      int64
ORIGIN                 object
DEST                   object
CRS_DEP_TIME           int64
DEP_TIME               float64
DEP_DELAY              float64
TAXI_OUT               float64
WHEELS_OFF             float64
WHEELS_ON              float64
TAXI_IN                float64
CRS_ARR_TIME           int64
ARR_TIME               float64
ARR_DELAY              float64
CANCELLED              float64
CANCELLATION_CODE      object
DIVERTED               float64
CRS_ELAPSED_TIME       float64
ACTUAL_ELAPSED_TIME    float64
AIR_TIME               float64
DISTANCE               float64
CARRIER_DELAY          float64
WEATHER_DELAY          float64
NAS_DELAY              float64
SECURITY_DELAY         float64
LATE_AIRCRAFT_DELAY    float64
Unnamed: 27            float64
```

Figure 2 Original Data Types

```
In [29]: for col in airline.columns:
    ...:     print("There are: %d null values in column %s" % (airline[col].isnull().sum(),col))
There are: 0 null values in column FL_DATE
There are: 0 null values in column OP_CARRIER
There are: 0 null values in column OP_CARRIER_FL_NUM
There are: 0 null values in column ORIGIN
There are: 0 null values in column DEST
There are: 0 null values in column CRS_DEP_TIME
There are: 112317 null values in column DEP_TIME
There are: 117234 null values in column DEP_DELAY
There are: 115830 null values in column TAXI_OUT
There are: 115829 null values in column WHEELS_OFF
There are: 119246 null values in column WHEELS_ON
There are: 119246 null values in column TAXI_IN
There are: 0 null values in column CRS_ARR_TIME
There are: 119245 null values in column ARR_TIME
There are: 137040 null values in column ARR_DELAY
There are: 0 null values in column CANCELLED
There are: 7096862 null values in column CANCELLATION_CODE
There are: 0 null values in column DIVERTED
There are: 10 null values in column CRS_ELAPSED_TIME
There are: 134442 null values in column ACTUAL_ELAPSED_TIME
There are: 134442 null values in column AIR_TIME
There are: 0 null values in column DISTANCE
There are: 5860736 null values in column CARRIER_DELAY
There are: 5860736 null values in column WEATHER_DELAY
There are: 5860736 null values in column NAS_DELAY
There are: 5860736 null values in column SECURITY_DELAY
There are: 5860736 null values in column LATE_AIRCRAFT_DELAY
There are: 7213446 null values in column Unnamed: 27
```

Figure 3 Columns with Null Values

We have renamed the columns and extracted information from the date column to get information like month, day date etc. we converted the data types of many variables since few of them were supposed to be used as numeric, but they were in object format. we have also numerically coded the columns for column cancellation_code

As you can see in the above screenshot there is an unnamed column, we removed the column.

```
array(['UA', 'AS', '9E', 'B6', 'EV', 'F9', 'G4', 'HA', 'MQ', 'NK', 'OH',
       'OO', 'VX', 'WN', 'YV', 'YX', 'AA', 'DL'], dtype=object)
```

For Linear Regression Model

We renamed the original data column names and validated the nulls, however with a little different approach. We first plotted a density plot for chosen attributes. After plotting the density plot with columns with "nan" values, we found none of the columns strictly follows normal distribution, and most of them were largely skewed and concentrated to only few values (see figure 4). Replacing methods, we tried included applying fillna () method to replace "nan", and replacing missing "nan" values with the mean of corresponding columns. However, none of the methods enable us to develop model with desirable results. So instead of replacing "nan" with normal distribution, we decided to use merely replace "nan" with extreme values that without the original data range.

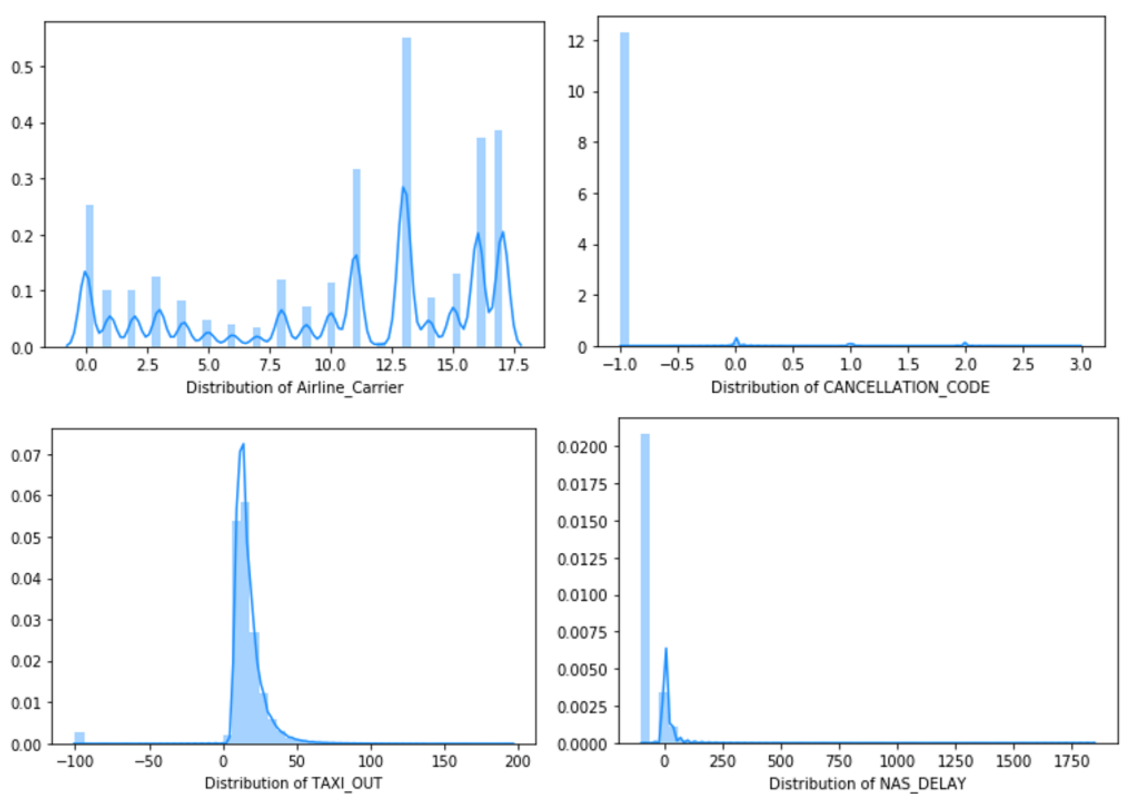
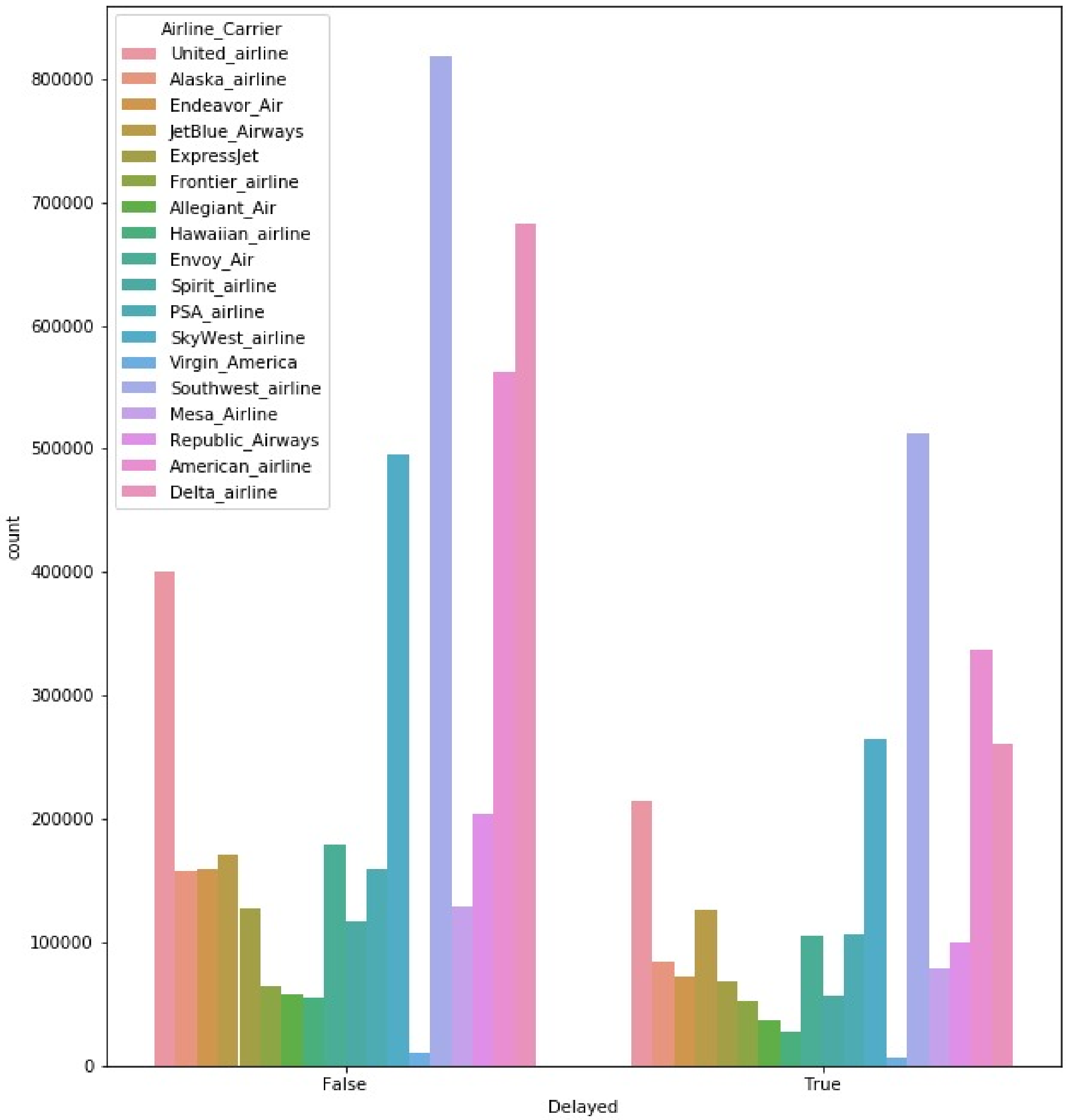


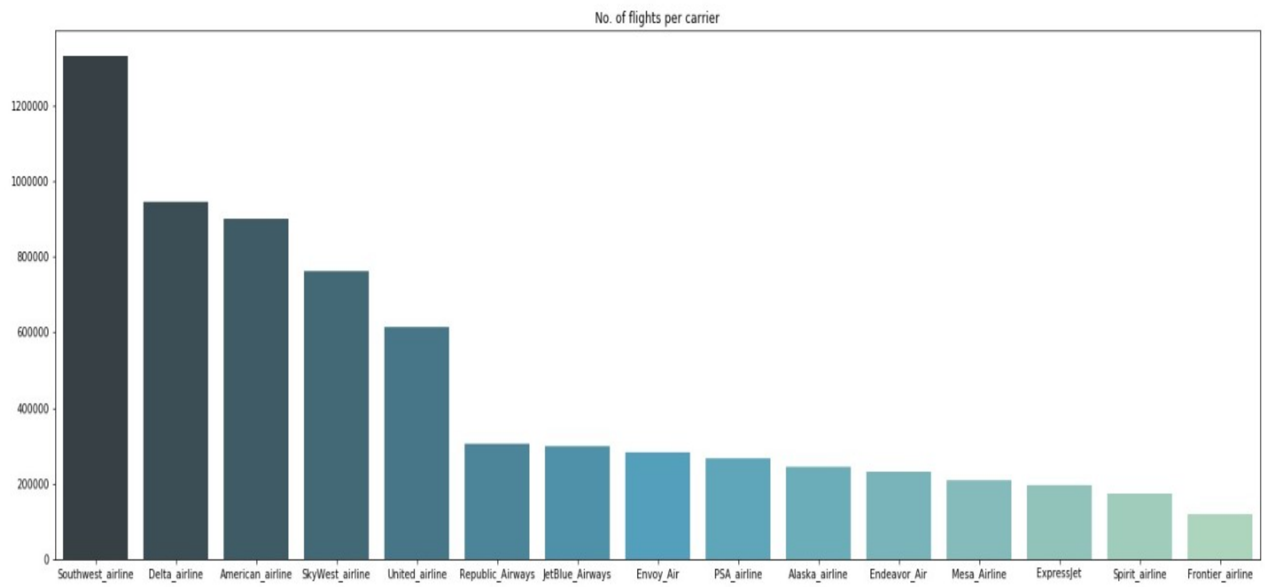
Figure 4 Four Examples of Data Density Plots (Distribution)

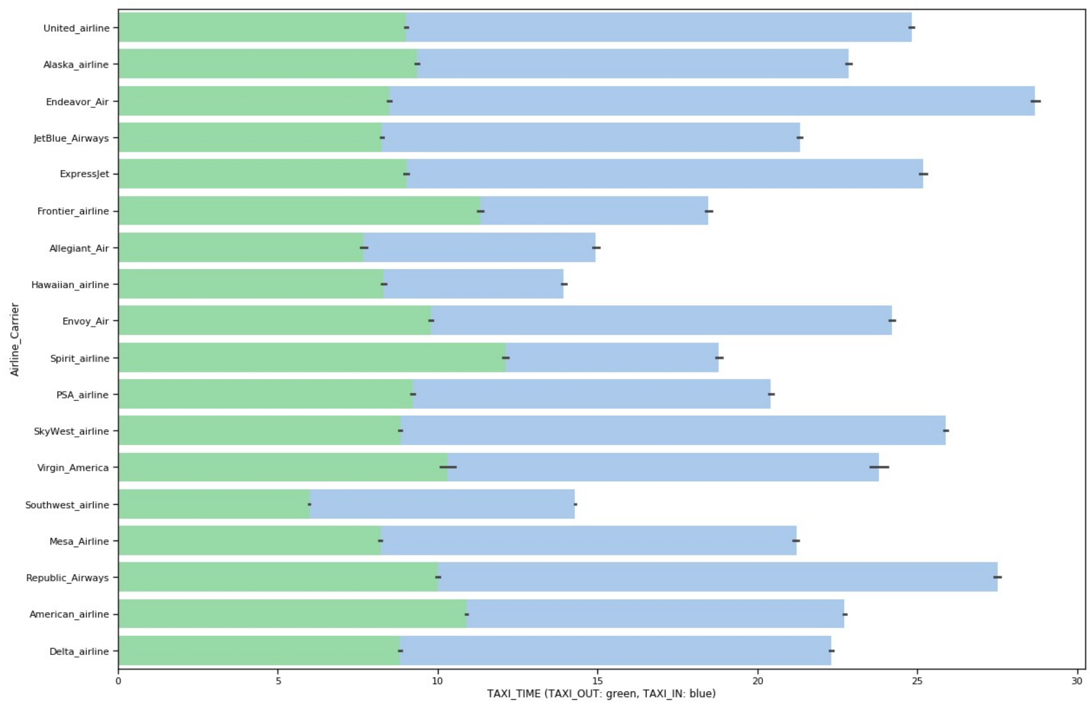
Initial Data Exploring

After data cleaning we start the first process of exploring our data if there were any patterns within the independent variables.



The above graph shows the no of delays airline wise. On the left side you can see there is a false value, which means instances when an airline has not been delayed. On the right-side true values suggests that there is a delay. We can see and conclude that maximum delay is caused by Southwest Airlines. Also, in the next graph we can see that the maximum number of flights are from Southwest Airlines, which compel us to think that one of the reasons for the delay is the operational process of airline. And these delays are known as career delay, we can reduce this delay with effective planning strategies.





As we can see the above graph Taxi in And Taxi out time are almost similar for most of the airlines. but there for endeavor airlines & republic airways, the taxi out time is much higher than taxi in time. We were not able to find out the exact cause for this, so we consider it an anomaly.

Dimensionality Reduction:

So as there were 28 columns and we wanted our model to be very precise, before going ahead we wanted to be sure there should not be any kind of correlation between the predictor variables, otherwise our model will be overfitting. So, we used correlation matrix and the criteria was, if 2 variables have correlation greater than 0.4 or less than -0.4, we will drop one of those variables.

The most common correlated variables are actual departure time, actual arrival time, planned arrival time planned departure time among others, this makes sense because these factors are directly impacting the delay so there is no point of adding those variables.

Model Building

Using Linear Regression to Predict how much time we should expect a delay

Since logistic regression is appropriate for categorical values, and we expect to predict the potential delayed time, which is a numerical valuable, it makes more sense to apply Linear Regression for our model. Therefore, we applied sklearn linear model, and used r2_score to evaluate our model. We set Delay_Departure_Time, which is a set of both positive and negative figures to imply how long exactly a plane departure delayed/early[2]. We then also include week, month, airline carrier reference, planned elapsed time (in air time), distance between two departure and destinations, flight planned departure time, departure airport code, and taxi-in and taxi-out[3] time.

We split the test-train sets into 2:8 ratio, and got a R Square score of 0.806, which was acceptable.

```
yDelay = airline["Delay_Time_Departure"]
xDelay = airline[["Month","Week","Airline_Carrier","TAXI_OUT","TAXI_IN","Planned_Elapsed_Time","DISTANCE","Planned_Departure_Time","Airport_Departure_Code"]]
```

Figure Dependent and Independent Variables

```
Linear model accuracy (with the test set): 0.8062686700094936
```

Figure 5 Linear Model Accuracy

Logistic Model : For logistic Model we used the following Dependent variables

Actual Arrival time - Expected Arrival Time + (Actual Departure time - Expected Departure Time. But this independent variable could be any random number so we created another column in which if total delay > 0 then value will be 1 else it will be 0. This made us think that we should run logistic regression on this model and predict the factors responsible for delay. For testing we use 25 % of the dataset and we used 75 % of dataset for training our model. After running our model, we found out that our model is 82 % accurate and below is the confusion matrix.

```
[101] confusion_matrix = pd.crosstab(y_test, y_pred,
      rownames=['Actual'], colnames=['Predicted'])...
```

Predicted	0	1
Actual		
0	946777	187123
1	179310	454745

Factors which are affecting the delay:

Airline_Carrier - (AA, UA, …)

Month - Which Month (Jan, Feb…)

Week - Which date of the week (Monday, Tuesday, etc.)

Planned_Elapsed_Time - Planed in-air time

Tax In/Tax Out

DISTANCE - that's straightforward

Planned_Departure_Time

Airport_Departure_Code - Which airport the plane is going to start

Conclusion

After applying both the models for predicting whether a flight should be delayed, as well as how much one would expect a flight should be delayed, we found the following factors to be important: week, month, airline carrier reference, planned elapsed time (in air time), distance between two departure and destinations, flight planned departure time, departure airport code, and taxi-in and taxi-out[4] time. By applying our model, on the data collected, one could be able to predict whether a flight might be delayed, and more importantly, how long delayed time she/he would expect.

However, there are some limitation in our model, first, our model only included one-year data due to our computation capability, as more years of data included, the prediction could be easier. In addition, some other related information such as airplane type, e.g., detailed weather data specific to airport[5] were not included. Therefore, researchers could try to collect more related data and deploy better computational powers to build a better model.

Taxi in/out time means the time when a flight wheel was on/off to the time the flight gate in/on time. See Aviation System Performance Metrics: https://aspmhelp.faa.gov/index.php/ASPM_Taxi_Times:_Definitions_of_Variables .
The only data related with weather is WEATHER_DELAY, which indicates whether a delay caused by weather. And there are also too many missing values in this attribute. Usually, a plane delayed would cause problem to passengers, as it might lead to missing the following airline in certain cases. However, a plane leave early would not cause too much problem, as airlines usually only depart early when it is sure that would not cause any troubles, in another word, usually only the airlines are sure the flight would depart early, would airline allow a flight to depart early. For more references, see: fox news reporter Rick Seaney article "Do flights ever leave early? And 4 other common travel questions",
https://www.foxnews.com/travel/do-flights-ever-leave-early-and-4-other-common-travel-questions
Taxi in/out time means the time when a flight wheel was on/off to the time the flight gate in/on time. See Aviation System Performance Metrics:
https://aspmhelp.faa.gov/index.php/ASPM_Taxi_Times:_Definitions_of_Variables .