Machine Learning Engineer Nanodegree

Predicting flight delays

Srividya
August 17th, 2019

I. Definition

Project Overview

The commutes through flights have helped us to overcome the barrier of geography.They have become very common among the individuals, might be a trip or a business meeting. As the number of people who travel have increased, the number of airlines and airports have also increased to make a passenger's life much easier. The sizes of the airports differs from one to another based on the population of the passengers and number of incoming and outgoing flights in the airports. There are different types of flights, in this problem we mainly focus on domestic flights. There are nearly 44K aircrafts that lands daily on average in US and 87K flights that crisscross US everyday on avg. From this we can infer that the Air traffic is huge and there will be delays in takeoffs and landings. The delay in flight is very painful for passengers and airlines. This could lead to very serious problem such as missing important business meetings or a sport competition etc which might lead to huge losses. This mainly creates stressful situations among the passengers. There are cases where passengers have missed connecting flights and have waited double the time than they expected.
Due to this problem it has become very important that passengers are well informed in advance so that they can plan their trip accordingly. If we have data on flights arrival and departure timings of its everyday trip we can capture the pattern of these delays and predict them accordingly based on the data(history) that we have. This is the main role of our project.

Problem Statement

Booking flights online have become very common through websites. There are tons of websites that do flight bookings, they provide details such as arrival, departure time, fare etc. But what if they provide details on the delays? This will help users or passengers to decide or plan on their trip upfront. The overhead of delays or surprise delays will be reduced to some extent.
This prediction also helps airlines to avoid delays through planning and thus help both passengers to have a good journey and airlines to become more efficient. We can predict these delays using regression models.

Metrics

As our problem is based on prediction its a regression problem and large differences between actual and predicted are punished more in MSE than in MAE. Therefore we have taken MSE as our evaluation metric. As we have tried different models including benchmark model which is linear regression they all should have common evaluation metrics so that we can compare which is the better model. For this purpose we are using MSE as a common metric. MSE is given by:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

MSE basically measures average squared error of our predictions. For each point, it calculates square difference between the predictions and the target and then average those values.

The higher this value, the worse the model is. It is never negative, since we're squaring the individual prediction-wise errors before summing them, but would be zero for a perfect model .

II. Analysis

Data Exploration

Each entry of the input file `flights.csv` file corresponds to a flight and we see that more than 5'800'000 flights have been recorded in 2015. These flights are described according to 31 variables. Here are few variables that I explain:
Inputs from the file:

- **YEAR, MONTH, DAY, DAY_OF_WEEK**: dates of the flight
- **AIRLINE**: An identification number assigned by US DOT to identify a unique airline
- **ORIGIN_AIRPORT** and **DESTINATION_AIRPORT**: code attributed by IATA to identify the airports
- **SCHEDULED_DEPARTURE** and **SCHEDULED_ARRIVAL** : scheduled times of take-off and landing
- **DEPARTURE_TIME** and **ARRIVAL_TIME**: real times at which take-off and landing took place
- **DEPARTURE_DELAY** and **ARRIVAL_DELAY**: difference (in minutes) between planned and real times

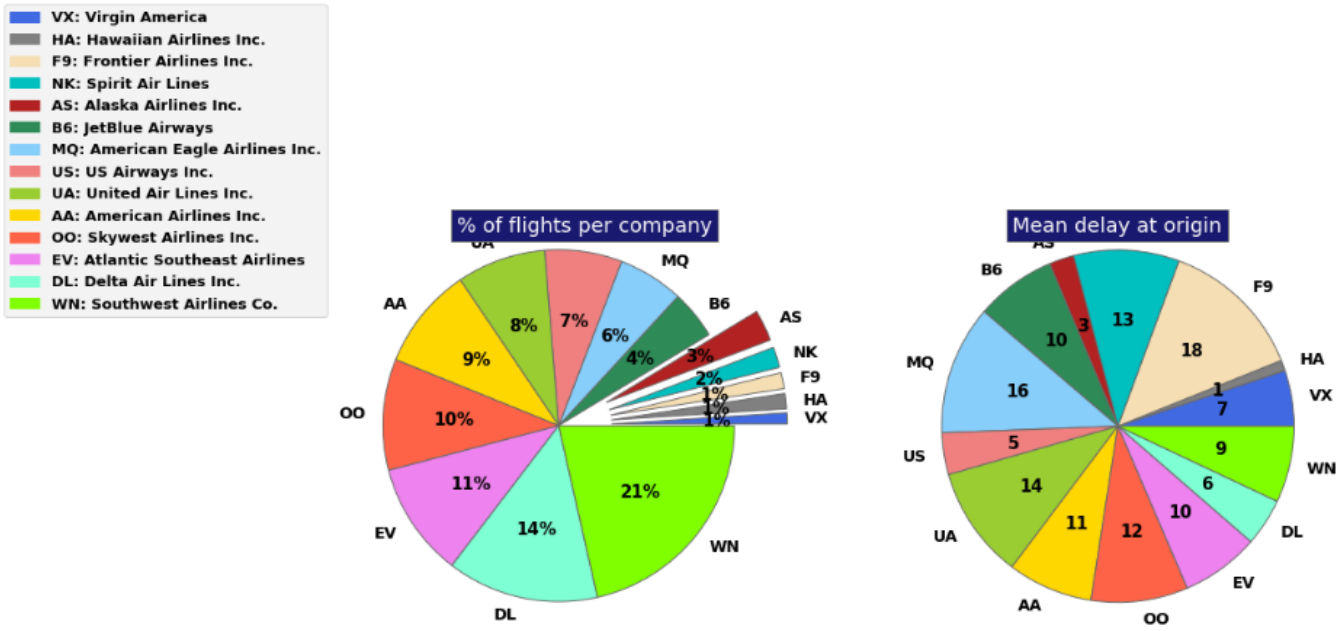
- **DISTANCE**: distance (in miles)
  Output:
  Y-Predicted delay values
  An additional file of this dataset, the `airports.csv` file, gives a more exhaustive description of the airports

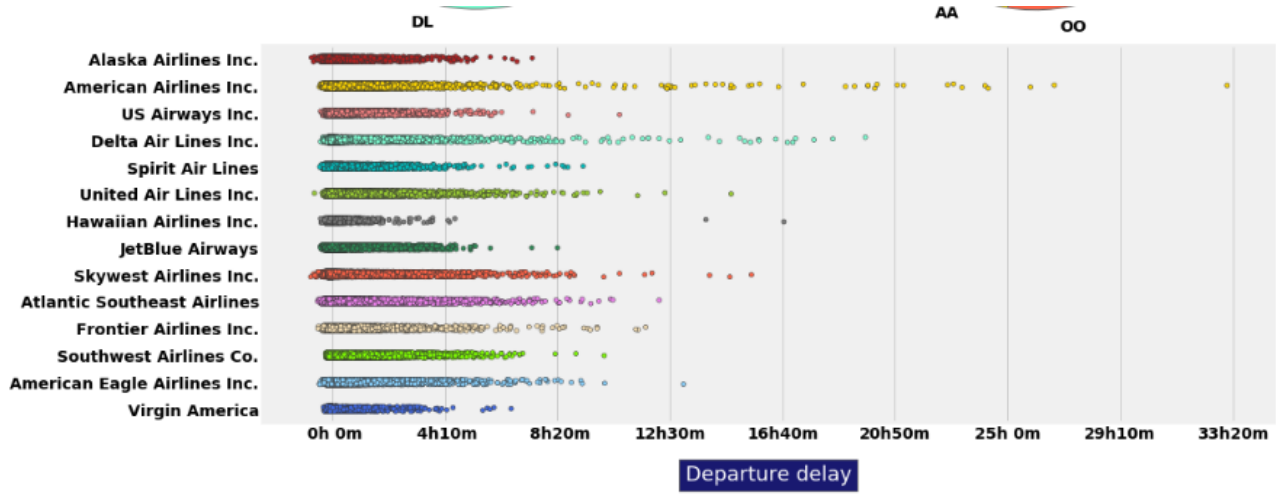
For easy computations I will be taking subset of this data which is of one month January 2015.

Next I will clean the data by changing the date time format of time variables for convenience. Also as there are few missing values but very less. I remove them and remove all NANs.
The statistics of the the delay grouped by each airline is as follows:

AIRLINE	count	max	mean	min
VX	4647.0	397.0	6.896277	-20.0
HA	6408.0	1003.0	1.311954	-26.0
F9	6735.0	696.0	17.910765	-32.0
NK	8632.0	557.0	13.073100	-28.0
AS	13151.0	444.0	3.072086	-47.0
B6	20482.0	500.0	9.988331	-27.0
MQ	27568.0	780.0	15.995865	-29.0
US	32478.0	638.0	5.175011	-26.0
UA	37363.0	886.0	13.885555	-40.0
AA	43074.0	1988.0	10.548335	-29.0
OO	46655.0	931.0	11.999957	-48.0
EV	48084.0	726.0	9.678895	-33.0
DL	63676.0	1184.0	5.888215	-26.0
WN	98060.0	604.0	9.453426	-15.0

Exploratory Visualization

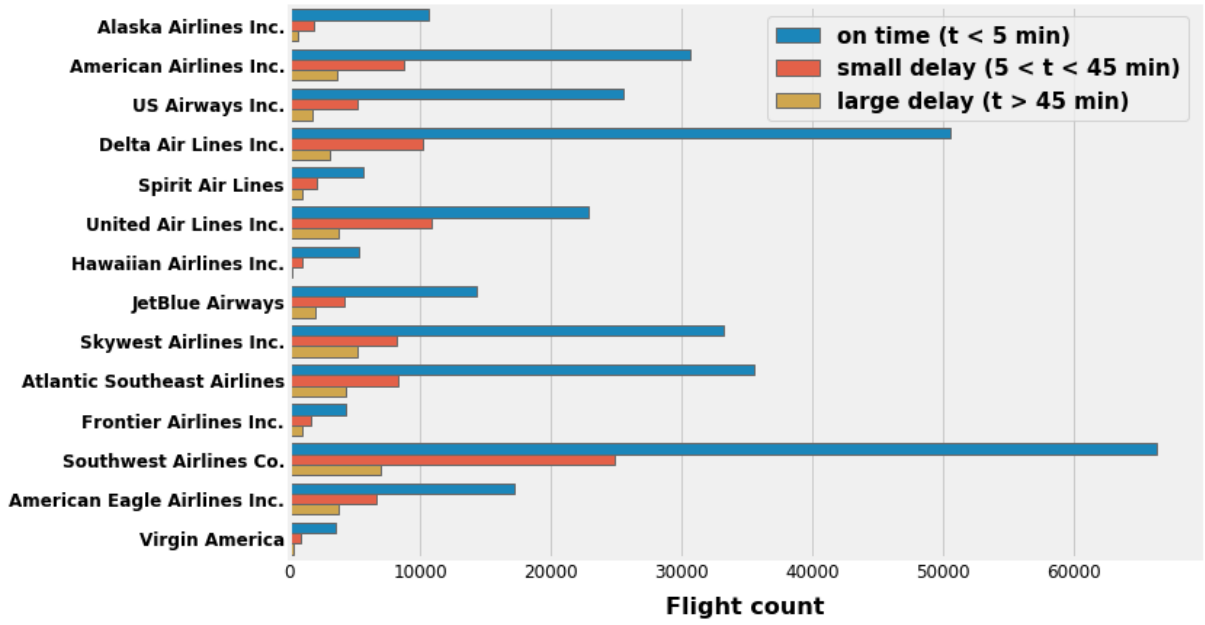




Considering the first pie chart that gives the percentage of flights per airline, we see that there is some disparity between the carriers. For exemple, *Southwest Airlines* accounts for ~20% of the flights which is similar to the number of flights chartered by the 7 tiniest airlines. However, if we have a look at the second pie chart, we see that here, on the contrary, the differences among airlines are less pronounced. Excluding *Hawaiian Airlines* and *Alaska Airlines* that report extremely low mean delays, we obtain that a value of 11+or-7 minutes would correctly represent all mean delays.

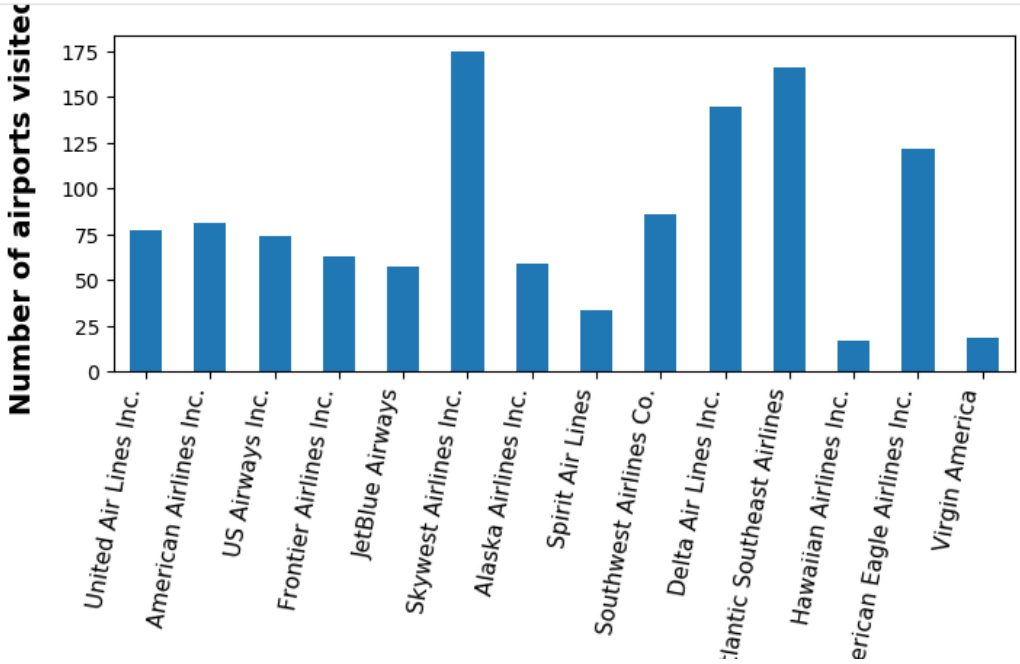
Finally, the figure at the bottom makes a census of all the delays that were measured in January 2015. This representation gives a feeling on the dispersion of data and put in perspective the relative homogeneity that appeared in the second pie chart. Indeed, we see that while all mean delays are around 10 minutes, this low value is a consequence of the fact that a majority of flights take off on time. However, we see that occasionally, we can face really large delays that can reach a few tens of hours !

The large majority of short delays is visible in the next figure:



This figure gives a count of the delays of less than 5 minutes, those in the range 5 < t < 45 min and finally, the delays greater than 45 minutes. Hence, we wee that independently of the airline, delays greater than 45 minutes only account for a few percents. However, the proportion of delays in these three groups depends on the airline: as an exemple, in the case of
SkyWest Airlines, the delays greater than 45 minutes are only lower by ~30% with respect to delays in the range 5 < t < 45 min. Things are better for *SoutWest Airlines* since delays greater than 45 minutes are 4 times less frequent than delays in the range 5 < t < 45 min.

Number of airports visited by each airline:



Algorithms and Techniques

Here we are using the following models:
-Linear regression
-Polynomial regression
-Light GBM

Here I will use Linear regression as benchmark model and later I move to polynomial regression as it provides of the relationship between the dependent and independent variable.Polynomial basically fits a wide range of curvature. Later we will use Light GBM as I wanted to try out a tree based model.
We split our data to train test and apply the above algorithms and compare the evaluation metrics.
The regularization technique ridge is also implemented for better outcome. We have also used Grid search technique to find best plynomial order n and alpha-coefficient of ridge regression.

Benchmark

As any regression problem I have considered linear regression as benchmark model and have tried to beat its MSE value
The below is the value of the MSE for our benchmark model that was used:
Model MSE
Linear Regression 76.95 (Benchmark)

Liner regression can be mathematically expressed as:
Y = mX+b where:
Y = Target variable or dependent variable, in our case it is delay
X=x1, x2 … xn = Independent variables (features) which will impact the Target variable. In our case, it can be source airport, airline, time of departure etc m = set of coefficients of different features b = Bias, it is also Y interecept

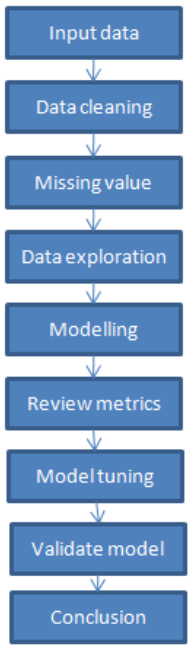
III. Methodology

Data Preprocessing

We have few data processing steps such as changing the format of Scheduled departure(Date_time), Scheduled arrival(time) and Arrival time(time) and departure time(time) to the formats specified in brackets. Scheduled departure was in float type.
Next we treat missing values by removing them as ~97% of our data is complete.
While doing analysis we do it for the month of january.
In modelling we split the data into train and test. Also we have considered cross validation and regularization.

Implementation

Here are the following steps that were followed while solving this problem:



Refinement

We will preprocess the data by changing the date time formats as required and treat missing values.
We split the data into train (70%) and test(30%). We include regularization technique and use grid search to find best polynomial order and alpha coefficient of ridge regression.
The grid search result were as below:

```
n=1 alpha=0 , MSE = 60.422
n=1 alpha=2 , MSE = 59.539
n=1 alpha=4 , MSE = 59.522
n=1 alpha=6 , MSE = 59.828
n=1 alpha=8 , MSE = 60.251
n=1 alpha=10 , MSE = 60.707
n=1 alpha=12 , MSE = 61.157
n=1 alpha=14 , MSE = 61.586
n=1 alpha=16 , MSE = 61.988
n=1 alpha=18 , MSE = 62.361
n=2 alpha=0 , MSE = 1698.4
n=2 alpha=2 , MSE = 59.986
n=2 alpha=4 , MSE = 59.776
n=2 alpha=6 , MSE = 59.757
n=2 alpha=8 , MSE = 59.818
n=2 alpha=10 , MSE = 59.927
n=2 alpha=12 , MSE = 60.066
n=2 alpha=14 , MSE = 60.225
n=2 alpha=16 , MSE = 60.395
n=2 alpha=18 , MSE = 60.573
```

IV. Results

Model Evaluation and Validation

The below table gives the MSE value obtained for used models

Model	MSE
Linear Regression	76.95
Light GBM	80
Polynomial regression	74.8

The Polynomial regression model is better than Linear regression with MSE ~74. Later I tried it for Light GBM which gave an MSE value of ~80

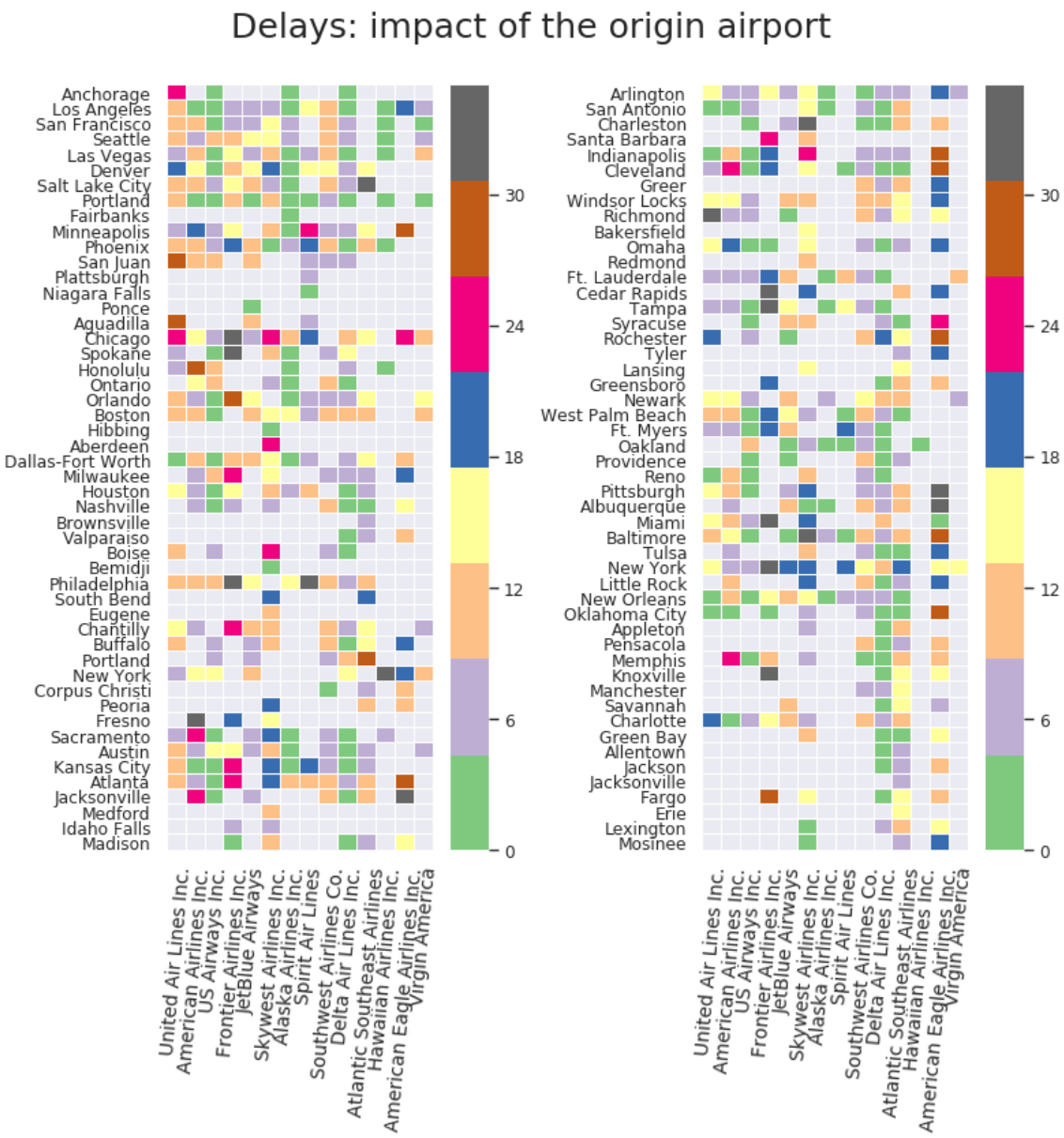
Justification

As the above figure gives the MSE value for each model, we can see that Polynomial regression is performing better than both the benchmark model and LightGBM model. This is due to the use of regulrization with parameter tuning.
There could be improvements made on the LightGBM as well by tuning their parameters but as we can see the params of LightGBM are very large and would take time for its tuning. This becomes a new project in itself.

V. Conclusion

Free-Form Visualization

Impact of origin airport on delays:



This figure allows to draw some conclusions. First, by looking at the data associated with the different airlines, we find the behavior we previously observed: for example, if we consider the right panel, it will be seen that the column associated with *American Eagle Airlines* mostly reports large delays, while the column associated with *Delta Airlines* is mainly associated with delays of less than 5 minutes. If we now look at the airports of origin, we will see that some airports favor late departures: see e.g. Denver, Chicago or New York. Conversely, other airports will mainly know on time departures such as Portland or Oakland.

Finally, we can deduce from these observations that there is a high variability in average delays, both between the different airports but also between the different airlines. This is important because it implies that in order to accurately model the delays, it will be necessary to adopt a model that is ** specific to the company and the home airport. This was very interesting analysis.

Reflection

We started of with basic linear regression and moved to polynomial regression with regularization. Later we have also used Grid search for this model. We have also used a lightGBM model traditional gradient boost.
The most time consuming part was data preprocessing. The cleansing and conversion of date formats to date time format. Next it was a bit tricky in regularization part and their parameter tuning. LightGBM model was a new learning and have made things easier.

Improvement

We can consider tree based ensembled method for better results. Also we can make improvements in lightGBM model and explore more regression trees. We can use the current LightGBM as bench mark or Polynomial regression and use the above mentioned models as improvements.