Modelling and Predicting the Flight Delays

Capstone Project| Udacity Machine Learning Engineer Nanodegree| Proposal

Karthik Sunil

Aug 8th, 2019

This document explains the background for this capstone project.

- The source of dataset
- Understanding the data and data cleaning
- Exploratory data analysis approach
- Different visualizations that can be used during data exploration
- Feature exploration that can be used for ML modelling
- Exploring different approaches

Domain Background

In the US, there are around 44,000 flights per day on an average. More than 2.5 million passengers travel through domestic flights per day. Air travel has become very stable and matured over couple of decades. However, there are still a lot of cancellations and delays that cause passengers a huge discomfort as well as loss in the businesses.

There are different type of people who use the commercial flights for travel- tourism, business trips or personal trips to work/home. Delays in the flights will have impact on:

- Connecting flights
- Missing the business meetings
- Customer discomfort

In this era of information, almost everyone takes informed decisions in their personal and business activities. To make decisions on the flight booking and planning more informed, prediction of delays/cancellation plays a mojor role

Problem Statement

Flight booking services today provide the data about the price comparisons, no. of hops, flight time etc, however, it will add to customer delight if those services can provide a hint about predicted delays/cancellations. It becomes a great boon for travellers if they know the predicted delay/cancellation of flights much in advance which helps them to plan ahead.

Predcition of delays will also help airlines to plan ahead to avoid them. Also sometimes they can reengineer their operative processes/ procedures which can reduce the flight delays.

Based on the historical data of flights, one can figure out what are all the parameters which affect the delay/cancellation. A regression model can be built to predict such delays in the flight.

Datasets and Inputs

The datasets considered for this analysis and model building is from Kaggle https://www.kaggle.com/usdot/flight-delays This dataset by DOT of USA contains all the flight data during 2015 by 14 different airlines. It has around 5.8 million flight data with airlines, source and destination airport, scheduled and actual deprature, arrival times, taxi-out/in data etc.

This is perfect data for this data project

Solution Statement

Some definitions:

- Departure time: It is when a plane leaves the gate
- Arrival time: It is when the plane pulls up to the gate
- Scheduled time: Planned time of either departure or arrival
- Actual time: Actual time when the flight has departed or arrived
- Delay: Difference between planned time and actual time, for for departure or arrival
- Taxi-out: It is defined as the time spent by a flight between its actual off-block time and actual take-off time.
- Taxi-in: It is defined as the time spent after flight has landed till it reaches the block for disembarkment.
- Air time: Period during which a particular aircraft remains airborne. Also called wheels-off to wheels-on time.

By looking above definitions, we observe many parameters with respect to times during a flight. Some of them are interdependent. For egs. the arrival time depends on departure time, taxi-out, taxi-in and air time.

In this project, we look for predecting the Arrival delay.

The best approach to predict the delay in the data is Regression. Following are some of the regression methods we can explore:

1. Linear Regression
2. Polynomial Regression
3. Regularization methods like Ridge or Lasso Regression

Liner regression can be mathematically expressed as:
`Y = m1x1 + m2x2 + m3x3 + .. + mnxn + b` where:
Y = Target variable, in our case it is departure delay
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 a1, a2, .. an = Coefficients of different features b = Bias, it is also Y interecept

Similarly in case of *Polynomial Regression* one or more of the features have the degree more than 1 in the equation.

In *Regularization methods*, coefficient values are penalized by adding them L1 Regularization or in their squares, L2 Regularization to the loss function

Benchmark Model

As per Federal Aviation Administration (*FAA*), General Aviation departure delays are 15 min or less. That can be considered as baseline for delays. With respect to building the model, we can consider source airport and airline are two feature variables and build a *Linear Regression model*. That can be considered as our benchmark model.

Evaluation Metrics

Some of common metrics considered for measurement of the regression model would be: 1.Mean Absolute Error (MAE) 2. MSE (Mean Squared Error) 3. R2 Score (Coefficient of determination)

In this solution design, we use the MSE as the preferred metric for evaluating the model

Project Design

The general sequence of the steps followed for the project:

1. Data acquaintance

In this step, data is understood by high level statistical analysis. Get familiarized with data columns, observe the statistcal information and completeness of the columns. I intend to use `pandas` libraries to explore the data and use methods like `describe` get more statistical information about data.

The data set's shape is as follows. It has around 5.8 millions of flight data with 31 features. One of them (most probably the *Arrival delay*) will be target variable and rest of 30 columns need to be analysed to see which of them impacts the target variable.

> Dataframe dimensions: (5819079, 31)

More exploration will be done to check the correlations between the columns and to get early intuitions on feature selections

We can observe that almost 98% of data is filled for most important columns. So, we can consider that dataset is pretty complete for analysis.

Filling Factor

2. Data Pre-Processing

Most importan step in this project is to check for formats of data. For instance we can observe that, in the data, dates are coded according to 4 variables: **YEAR, MONTH, DAY**, and **DAY_OF_WEEK**. In fact, python offers the ***datetime*** format which is really convenient to work with dates and times and I plan convert the dates in that format.

Another aspect of preprocessing of data is encoding the categorical variables. We have Airline and Airports information in each flight, they are non-numerical categorical variables. We can use either Label Encoding OR One Hot Encoding

3. EDA - Exploratory Data Analysis

Once we have general acquaintance of data, we next use different visualization techniques using `seaborn` and `matplotlib` for conducting Exploatory Data Analysis. Different graphics can be plotted to visualize the data. This will help us in shortlisting the features those will have highest impact on our target variable.

The visualization will also help us in shortlisting different algorithms which can be used for Regression. We will see if there are any Linear relationship between target variable, if not we should look for some patterns from which we can get some polynomial relations.

Some of the graphics which can be used for EDA would be - Scatter plots - Histograms - Heatmaps - Bar and Pie charts

4. Feature Engineering

In this step we find the relevant featues necessary for building the model. We might also engineer new features based on other features using the techniches like PCA if feasible.

5. Model Selection

In this step, we experiment with existing algorithms to find out the best model which helps to resolve the problem at hand

6. Model Tuning

Once we have finalized the model, we can explore the tweaking of hyper parameters and other settings to improve the results

7. Testing

We need to split the data into training set and testing set. This is must to avoid any bias during the training. Step no. 5 thru 7 happens in loop. Every time we tweak something in algorithm and/or hyper parameters we use the train set to train and test set to evaluate the result using the prescribed metrics. This iteration of model improvements can be stopped once we have satisfactory results.

References:

- Predicting Taxi-Out Time at Congested Airports with Optimization-Based Support Vector Regression Method
- Flight Delay Information - Air Traffic Control System Command Center