Machine Learning Engineer Nanodegree

Capstone Proposal

Srividya
August 12th, 2019

Domain Background

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.

Datasets and Inputs

The datasets to be used for this analysis and prediction 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 columns like airlines, source and destination airport, scheduled and actual deprature, arrival times, taxi-out/in data etc. This perfectly suits the nature of the problem that we want to deal with in this project.

Solution Statement

Here are some definitions of the inputs that will be available: - Departure time: It is when the plane leaves the gate - Arrival time: It is when the plane pulls up to the gate - Scheduled time: Planned or expected 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.

As our project is on predicting delays which deals with the time most of the inputs are time based. Some are dependent and some are independent. For example the arrival time depends on departure time, taxi-out, taxi-in and air time.

In this project, we look for predicting the 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 = 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

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

The agency (FAA) categorises it as a delay if the flight takes more than 15 minutes.That can be considered as baseline for delays. With respect to building the model, we can consider source airport and airline as two variables and build a *Linear Regression model*. That can be considered as our benchmark model.

Evaluation Metrics

Most common metrics for a regression model are: 1. Mean Absolute Error (MAE) 2. MSE (Mean Squared Error) 3. R2 Score (Coefficient of determination)

In our solution we prefer mean squared error as our metric for evaluation.

Project Design

The flow or the sequesnce of this project is as follows:

1. Data acquaintance

In this step, we will have a high level statistical overview of the data by looking at the summary of the overall data. We get familiarised with the columns and the count of rows and columns. This also includes examining what type of data is each column so that later when we manipulate the data this would be helpful. As in any analysis I use `pandas, numpy` libraries to explore the data and use methods like `describe` get the summary of data and `shape` to get their counts. The dimension of the data is as follows: (5819079, 31) It has approximately 5.8M records with 31 columns among which are there are few features that would influence our delay time. Further one of the approach to find this would be finding out correlations between variables and so on. Missing values would also be a major part but as our data 98% of it is filled we can assume that our data is completely available for the analysis.

2. Data Pre-Processing

In this step we clean the data so that our model does not face unexpected consequences such has being biased etc. The data types of the columns can be changed if required for our convinience. For eg. The date format can be changed to any format convinient to us while doing EDA The other aspect could be converting categorical data to numerical data through onehot encoding such as airlines or airport names. This step also includes filling up missing values, but as stated earlier in this case the data is almost complete (~98%) therefore we can move on with other steps.

3. EDA - Exploratory Data Analysis

After having a higher level overview of the data we can move on to the deeper level of analysis for a better understanding of the data. This can be done by plotting graphs and getting useful insights out of it which would help us in building our model. We will use data visualization libraries such as `seaborn` and `matplotlib`. Data visualization help to find correlation between different variables and hence find interdependent variables. Some of the graphics which can be used for EDA would be - Scatter plots - Histograms - Heatmaps - Bar and Pie charts

4. Feature Engineering

As stated earlier the features will be engineered base on our requirements which included imputation (missing value treatment), onehot encoding etc., we might also engineer new features PCA if required.

5. Model Selection

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

6. Model Tuning

Once we have finalized the model, we can explore the tweaking of hyper parameters and other settings to improve the results. This might include grid search to find the best hyp-parameters for our model.

7. Testing

We will split the data into train test set also cross validation set to avoid overfitting which will avoid any bias during the training. We iterate from step 5 to 7. 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 evaluation metrics. We will stop this iteration once we have satisfying 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