1/17/2021

Flight Delay Prediction

Balavignesh G

iNeuron internship

**Flow Chart:**

Problem Statement

Data Collection from Kaggle

Store the data in MongoDB

Get the data from MongoDB to Jupyter Notebook

Exploratory Data Analysis

Feature Engineering

Feature Scaling

Split the data for Train, Test – 70/30

Model building using Decision Tree

Deploy the model in cloud

**Problem Statement:**

Finding and measuring factors affecting aircraft delays on the ground and in the air and developing machine learning algorithms to optimize airline and airport operations based on the factors responsible for the flight delay.

**Description:**

Air travel has been increasingly preferred among travellers, mainly because of its speed and in some cases comfort. This has led to phenomenal growth in the air traffic and on the ground. Increase in air traffic growth has also resulted in massive levels of aircraft delays on the ground and in the air.

There are multiple stages of operation of an aircraft taking place at terminal boundaries, runways, airports, and distinguished airspaces that may be susceptible to different kinds of delays or cancellations. Weather conditions, ground delays, air traffic control and several other constraints and unforeseen circumstances lead to delays and cancellations in the entire aviation industry.

Airlines have to bear high costs because of these delays and cancellations that include expenses on maintenance and compensations to travellers stuck in airports. It can even cause environmental harm by the rise in fuel consumption and gas emissions. Hence, it becomes an absolute necessity in the airline industry today to develop a machine learning solution that can accurately predict the delays in advance so that required efforts can be channelized towards maintenance of the aircrafts and in avoiding these delay situations by optimizing the flight operations.

**Data collection:**

For collecting the data, we approached the secondary data collection method from Kaggle which contains a dataset of Airlines, Airports, and Flights.

Link: [**https://www.kaggle.com/usdot/flight-delays?select=flights.csv**](https://www.kaggle.com/usdot/flight-delays?select=flights.csv)

**Data description:**

For flight delay prediction, we used 14 different Airlines from 321 Airports for United States domestic air-traffic of the year 2015 and months January, February, March. The dataset includes features as follows:

* Information about the flight (day, day of the week, airline, flight number, tail number)
* Information about origin and destination (origin airport, destination airport)
* Information about the departure (scheduled departure, departure time, departure delay, taxi-out, wheels-off)
* Information about the flight-journey (scheduled time, elapsed time, air time, distance)
* Information about the arrival (wheels-on, taxi-in, scheduled arrival, arrival time, arrival delay)
* Information about diversion, cancellation, and reason for the delay (air system delay, security delay, airline delay, late aircraft delay, weather delay)

**Dataset size:**

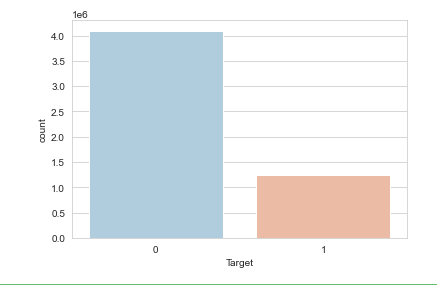
The size of the dataset contains 30 feature columns excluding the year as it is constant and 10,48,575 rows.

File size of Airlines: 359 Bytes

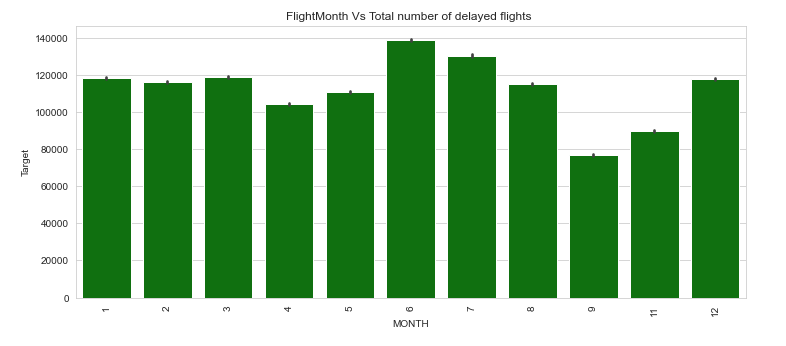
File size of Airports: 23.3 KB

File size of Flights: 564 MB

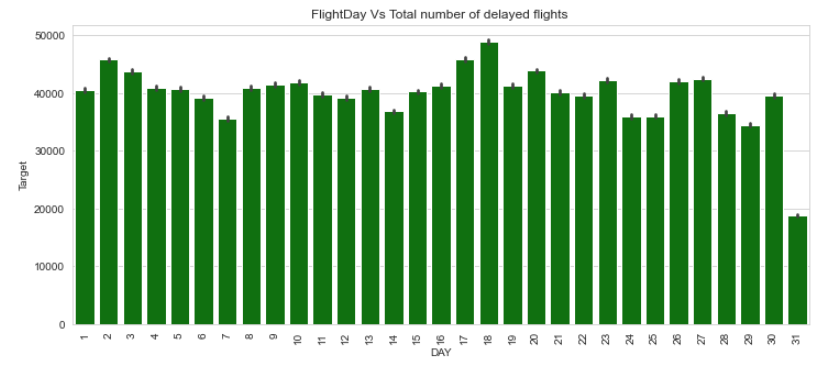
**Exploratory Data Analysis:**



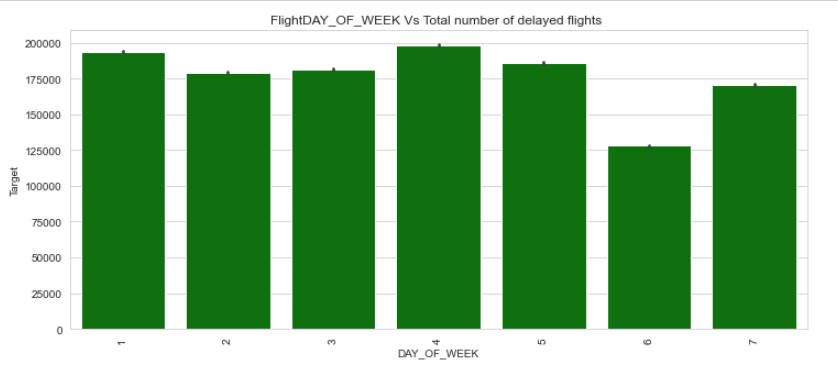
Flight delay is imbalanced, so let’s analysis when we have maximum no. of delays.



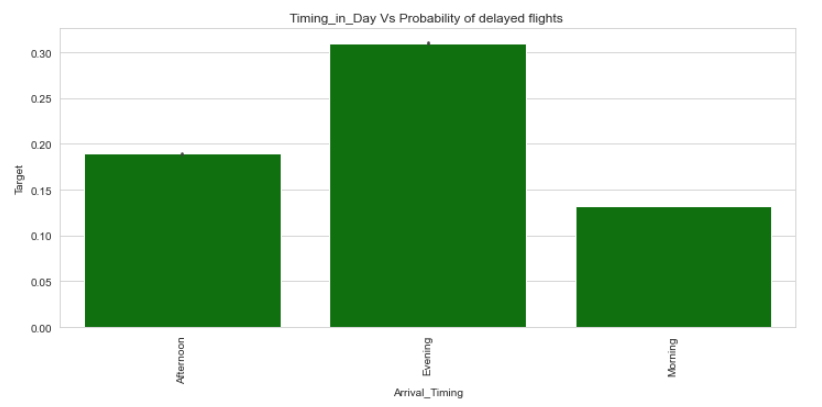
We experience maximum no. of delays during June & July of 2015.



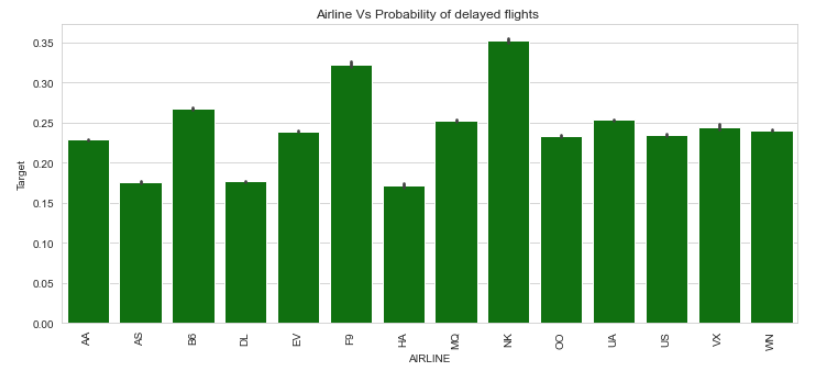
Every beginning and middle date of month experiences maximum no. of delays. At the end of the month we experience minimum delays.



Only on Saturday or may be on Weekends, delay in lesser. Monday & Wednesday the delay is high.



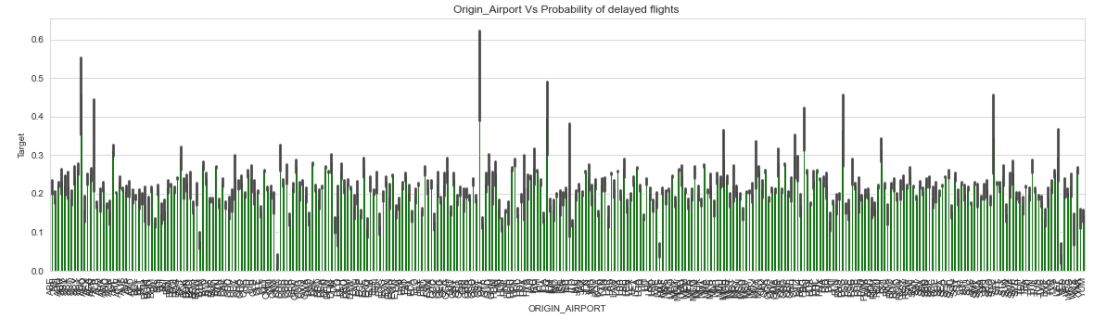
At Evening of everyday flight is delayed at the probability of 0.30 compare to Afternoon.

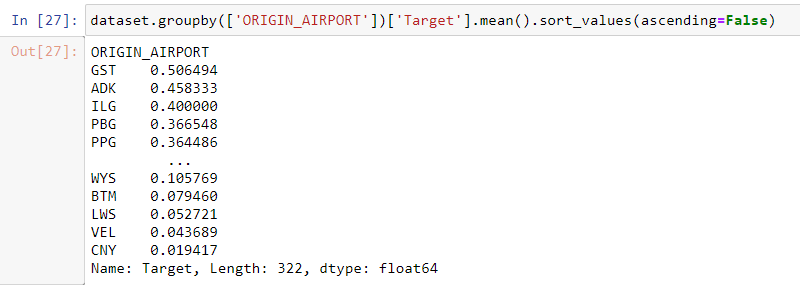


NK airline gets max no. of flight delays, p=0.35

F9 airline is in the second ordinal, p=0.33

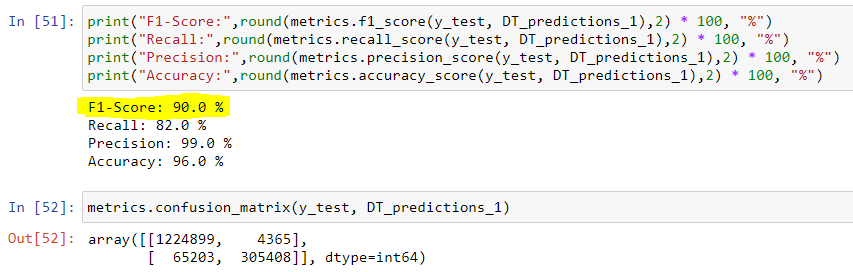
AS, DL, HA has min no. of flight delays, p=0.17





As analysed the top 5 airports have max no. of delays are GST, ADK, ILG, PBG, PPG.

**Model Building using Decision Tree:**



By Decision Tree algorithm and taking the features:

Day of Week, Origin Airport, Departure Delay, Air System Delay, Security Delay, Airline Delay, Weather Delay, Target

we get the **F1 Score: 90.0%**