CSE 519 Progress Report

Objective:

To build a 2 step model for predicting flight delays in minutes. The first model predicts whether there is a delay for a flight from an origin airport to a destination. If there is a delay, the second model predicts the expected delay in minutes for that specific flight.

Introduction:

The data for the project was taken from the U.S. Department of Transportation's (DOT) Bureau of Transportation which gives the delay details for each and every domestic flight between origin and destination airports. The data has various fields which are described under the dataset heading

Experimental setup:

Requires:

Hardware:

Minimum 4GB of RAM (GPU Optional)

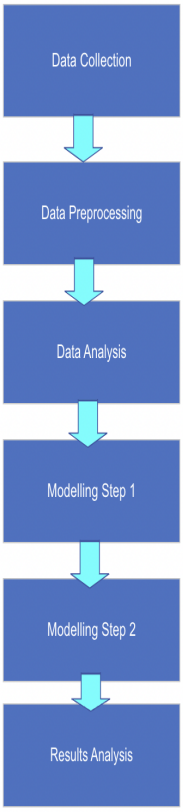
Software:

Anaconda Jupyter Notebook(Requires some pre installations)

Instead:

Google Colab

Project Flow Diagram



Dataset:

The dataset consists of data from 2009 to 2018. Size 8GB.

There are 3 types of fields that existed in the data:

A. Planned Fields: The fields which are estimated before the journey.
   1. 'FL_DATE'- Date of the flight
   2. 'OP_CARRIER'- Airline Identifier
   3. 'OP_CARRIER_FL_NUM'- Flight Number
   4. 'ORIGIN'- Starting Airport Code
   5. 'DEST'- Destination Airport Code
   6. 'DEP_TIME'- Actual Departure Time
   7. 'CRS_ARR_TIME'- Planned arrival time
   8. 'CRS_ELAPSED_TIME'- Planned time amount needed for the flight trip
   9. 'DISTANCE'- Distance between two airports

B. Calculated Fields: The fields which are calculated after the flight journey.
   1. 'CRS_DEP_TIME'- Planned Departure Time
   2. 'TAXI_OUT'- The time duration elapsed between departure from the origin airport gate and wheels off
   3. 'TAXI_IN'- The time duration elapsed between wheels-on and gate arrival at the destination airport
   4. 'WHEELS_OFF'- The time point that the aircraft's wheels leave the ground
   5. 'WHEELS_ON'- The time point that the aircraft's wheels touch on the ground
   6. 'ARR_TIME'- Actual Arrival Time
   7. 'CANCELLED'- Flight Cancelled (1 = canceled)
   8. 'CANCELLATION_CODE'- Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
   9. 'DIVERTED'- Aircraft landed on the airport that out of schedule
   10. 'AIR_TIME'- The time duration between wheels_off and wheels_on time
   11. 'CARRIER_DELAY'- Delay caused by the airline in minutes

12. 'WEATHER_DELAY'- Delay caused by weather
13. 'NAS_DELAY'- Delay caused by air system
14. 'SECURITY_DELAY'- Delay caused by security
15. 'LATE_AIRCRAFT_DELAY' - Delay caused by aircraft

C. Derived Fields: The fields which are derived from other fields

1. 'DEP_DELAY'- Total Delay on Departure in minutes.
2. 'ARR_DELAY'- Total Delay on Arrival in minutes
3. 'ACTUAL_ELAPSED_TIME'- AIR_TIME+TAXI_IN+TAXI_OUT

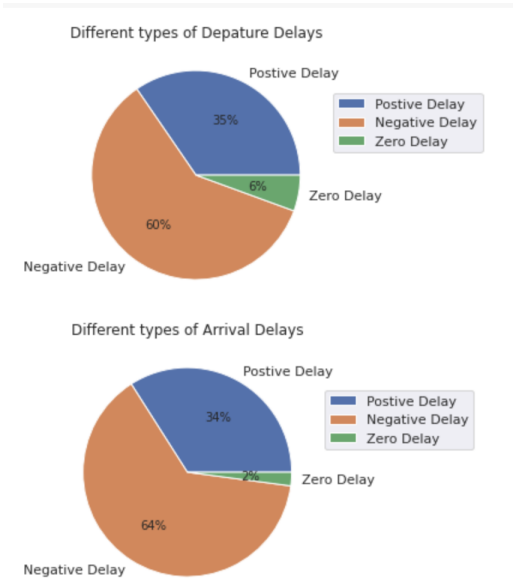
Problems with Data:

Challenges:

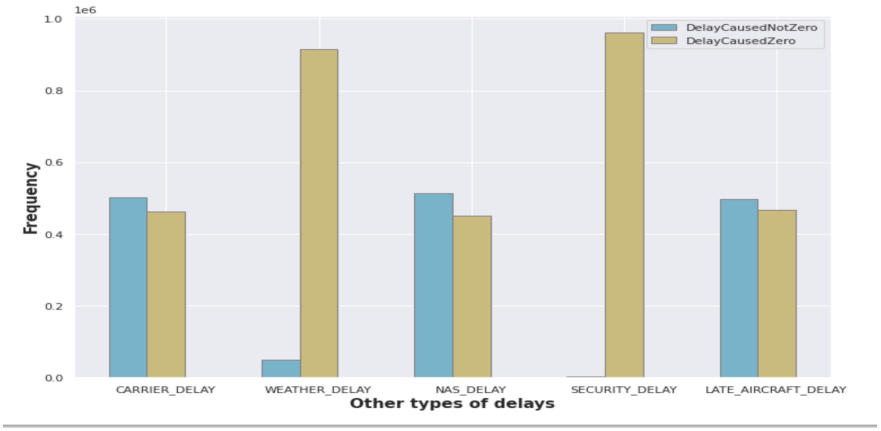
1. Extracting the data from the websites
2. Selecting the right columns for the data
3. Loading all of the data into the NoteBook(because data is huge It is around 8GB)

Data Analysis:

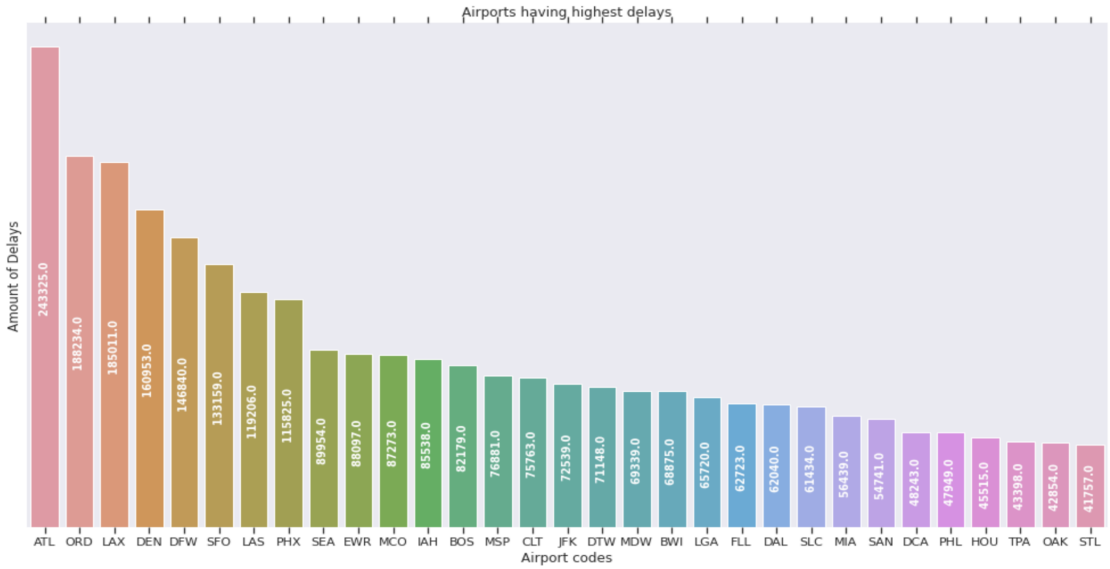
1. Plotted the pie graph to analyze the number of positive,negative and zero delays for both Departure and Arrival delay. Found out that there are more negative delays than the rest of the delay. Ie: There are more flights which arrive and leave earlier than the expected time.



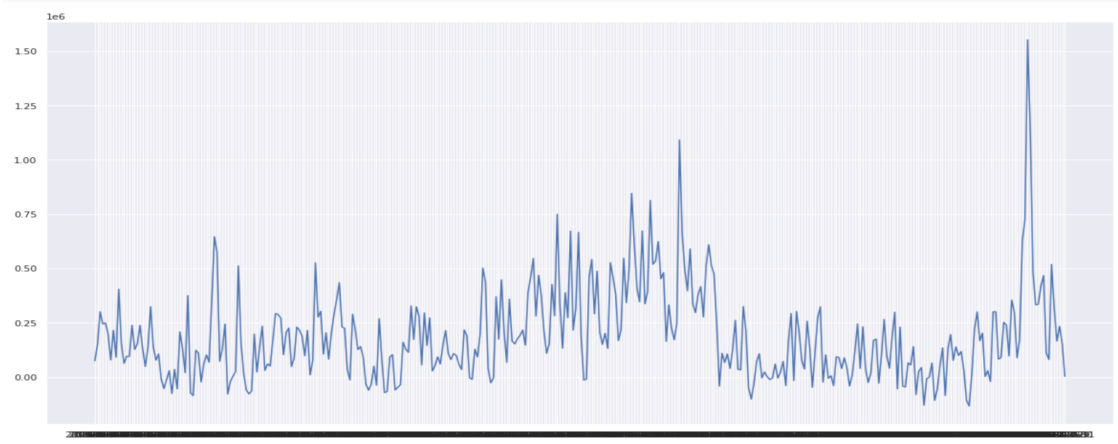
2. Plotted the bar graph to analyze the number of zero and non zero delay caused by the 'CARRIER_DELAY', 'WEATHER_DELAY' ,'NAS_DELAY', 'SECURITY_DELAY' , 'LATE_AIRCRAFT_DELAY'It shows that for some type of delays are caused more frequently.



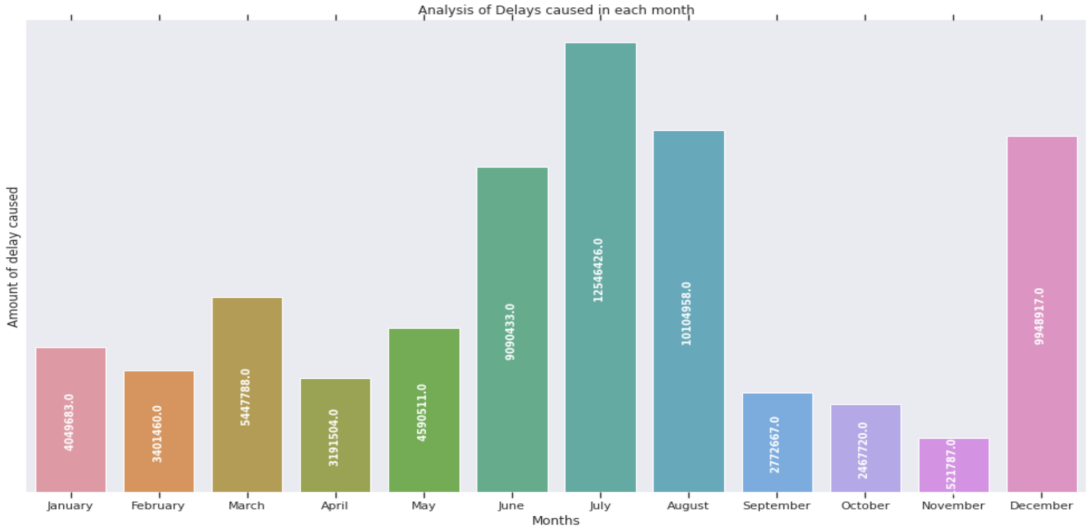
3. Plotted the graph of delay at each and every airport. Top 30 airports were shown below. It shows that there were few airports which had high chances of delay.



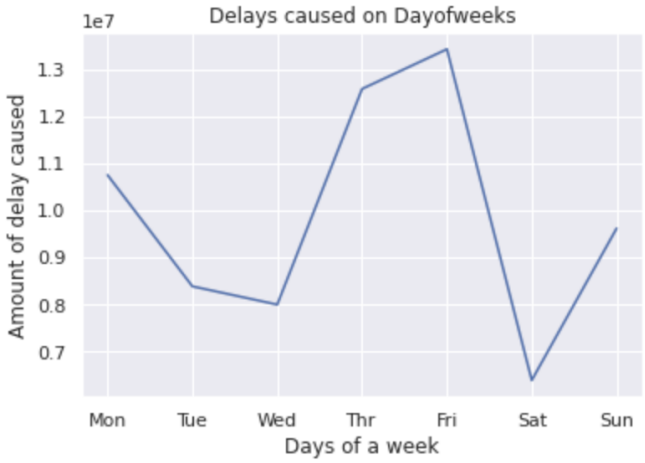
4. Plotted the graph of delay caused in each day of the year for a complete year. It shows that there were some days which had a high chance of delay.



5. Plotted the graph of delays caused in a month of the year. It shows that there were mid-months which have a high chance of delay.

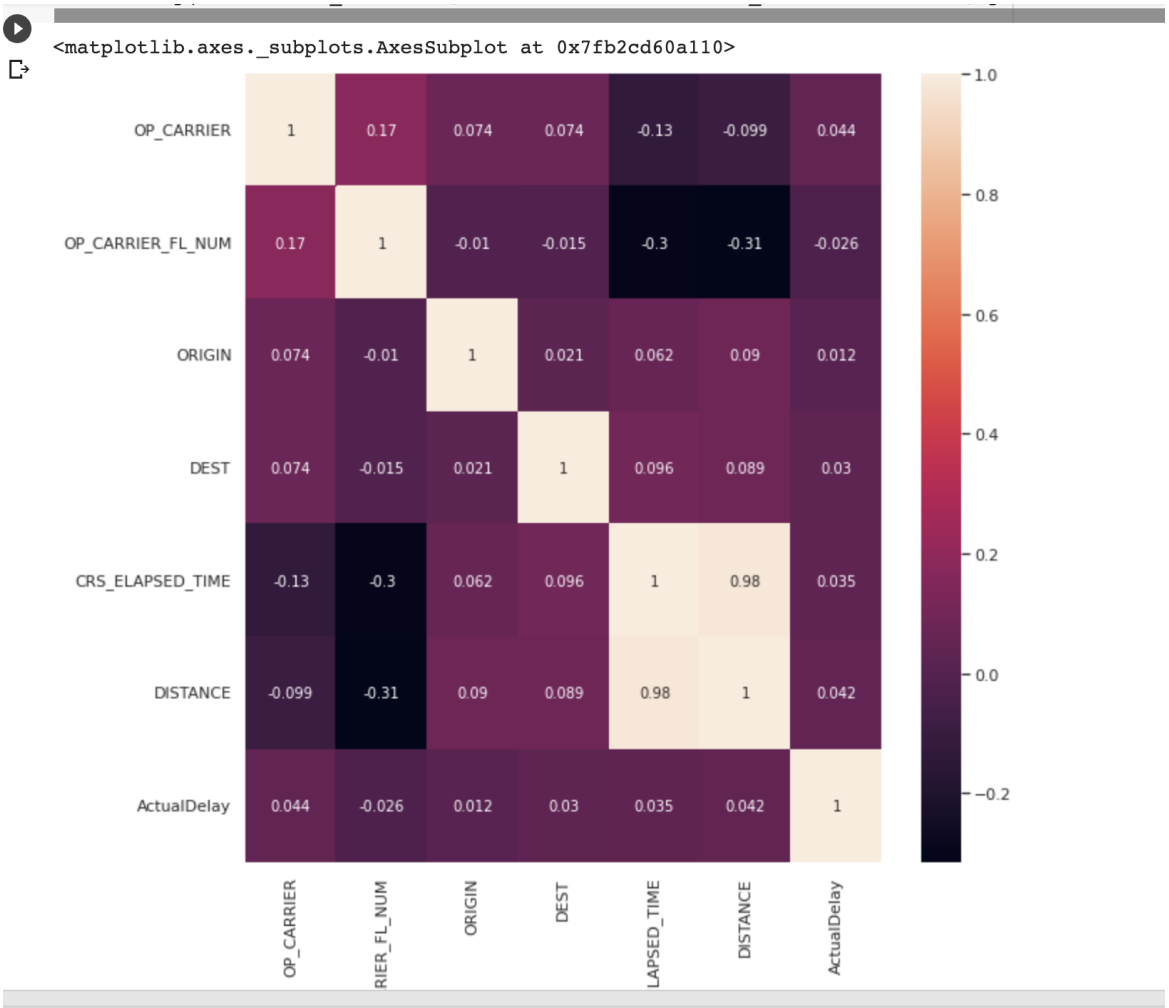


6. Plotted the graph of delay caused in every day of the week. It shows that there were Thursday and Friday has high chances of delay and has lower delay chances on Saturday.



Data preprocessing:

1. Filled the empty cells with 0
   Reason: We are considering the empty cell as value 0. As it is not affecting the delay.
2. Actual column was created which shows 1 when there is positive delay when we add the arrival and departure delay else 0
3. Plotted the heat map between multiple fields
   OP_CARRIER , OP_CARRIER_FL_NUM, ORIGIN, DEST, CRS_ELAPSED_TIME, DISTANCE, ActualDelay



4. The reasons for choosing the
   “OP_CARRIER, OP_CARRIER_FL_NUM, ORIGIN, DEST, CRS_ELAPSED_TIME, DISTANCE, ActualDelay”

1. From the above data analysis
2. From the correlation matrix

5. The OP_CARRIER, ORIGIN, DEST, ActualDelay were preprocessed by using Label encoding method

Implementation of Baseline:

1. Took the preprocessed data from the previous steps and divided into 80-20 ratio where 80% of data is used for training, rest of the data is used for testing
2. We have used Linear regression as the baseline model which results below

	precision	recall	f1-score	support
0	0.66	1.00	0.79	738386
1	0.33	0.00	0.00	385146
accuracy			0.66	1123532
macro avg	0.49	0.50	0.40	1123532
weighted avg	0.54	0.66	0.52	1123532

3. We have also used the XGBRegressor classification results which are nearly equal to the linear regression which is giving RMSE value 0.6476781459132928.

Next steps:

1. Currently, we had analysis from the 2016 year data. Similarly, we extend this to the rest of the data from 2009 to 2018 years.
2. Currently, we are taking weather data from BTS but we will take data from weatherAPIs to get a detailed analysis of which weather condition causes delays or cancellations.
3. We will make better predicting model by trying different algorithms and tune it to the better accuracy
4. We will build step 2 model to predict the delay time in minutes If step 1 model predicts there is a delay.
5. Also uses various statistical methods to extract more out of results.

References:

[1] M. Baluch, T. Bergstra, and M. El-Hajj, “Complex analysis of united states flight data using a data mining approach,” in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1–6, IEEE, 2017.

[2] Y. Ding, “Predicting flight delay based on multiple linear regression,” in *IOP conference series: Earth and environmental science*, vol. 81, p. 012198, IOP Publishing, 2017.

[3] J. J. Rebollo and H. Balakrishnan, “Characterization and prediction of air traffic delays,” *Trans portation research part C: Emerging technologies*, vol. 44, pp. 231–241, 2014.

[4] C. M. Ariyawansa and A. C. Aponso, “Review on state of art data mining and machine learn ing techniques for intelligent airport systems,” in *2016 2nd International Conference on Informa tion Management (ICIM)*, pp. 134–138, IEEE, 2016.

[5] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, “A review on flight delay prediction,” *arXiv preprint arXiv:1703.06118*, 2017.