Team 5 - Capstone - Summary Report

By: Fucheng Yao, Limei Huang, Eman Nagib, Kwangwoo Kim, Huaiping Wang
Github: https://github.com/enagib/capstone-2020

I. Introduction: Problem statement and setting

Problem Objective: "Using historical data, identify which model best predicts future flight delays and in turn identify top indicators of departure and arrival delays."

With the top U.S. airlines current high congestion rates and ever-increasing passenger traffic, the frustrations that come with unforeseeable delays are undeniable. Flight delays stem from several issues including carrier complications, late aircraft arrival, unfavourable weather, security threats, National Aviation System congestion, and other issues that might arise. Since these issues are not necessarily avoidable, the potential for predicting delays with respect to a specific airline, route, and date remains invaluable.

A few websites exist, such as Google Flights, that have recently started to provide some insight concerning potential flight delays. Google Flights displays these insights during the ticket purchase process, by showing 30-minute delay estimations as a precaution for their customers. The following project aims to take delay predictions a step further and leverage insight from historical data and machine learning algorithms to allow customers to see a refined delay prediction in respect to their future flight.

II. Dataset Background

Data set source: https://www.bts.gov/topics/airlines-and-airports-0

- We are using the Bureau of Transportation Airport - Airline and Airports. The BTS' Office of Airline Information publishes regular monthly and quarterly reports —on airline performance in the United States. Topics include airline origins and destinations, passenger traffic, on-time performance, holiday dates, etc.
- For this specific project, we have focused solely on data related to airline origins and destinations, on-time performance, delay reasons, and cancellation rates.

Weather data source: https://www.ncdc.noaa.gov/cdo-web/datatools/records;

- The National Oceanic and Atmospheric Administration(NOAA) has provided free access to global daily weather data since 1929. The dataset includes variables such as station id, year, month, average daily temperature, sea level pressure, visibility, wind speed, maximum gust, and weather indicators(fog, rain drizzle, snow, hail, thunderstorms, etc.). For the purpose of our project, we focused solely on stations within the US from the year 2015 to the year 2019.

Data cleaning and preparation:

Our data spans five years from 2015-2019 and includes only the top ten US domestic flights: United, American, Delta, Spirit, Southwest, Jet Blue, Hawaiian, frontier, Allegiant, and Alaska airlines. The dataset consists of flight identifying variables as well as route details, and weather condition variables. To

attain this data set we had to compile and merge US domestic flight data and national weather data for the aforementioned time span. The flight date and the airport code were used to merge the two datasets. Next we added variables to indicate what national holiday it was if a flight's date was on a national holiday date.

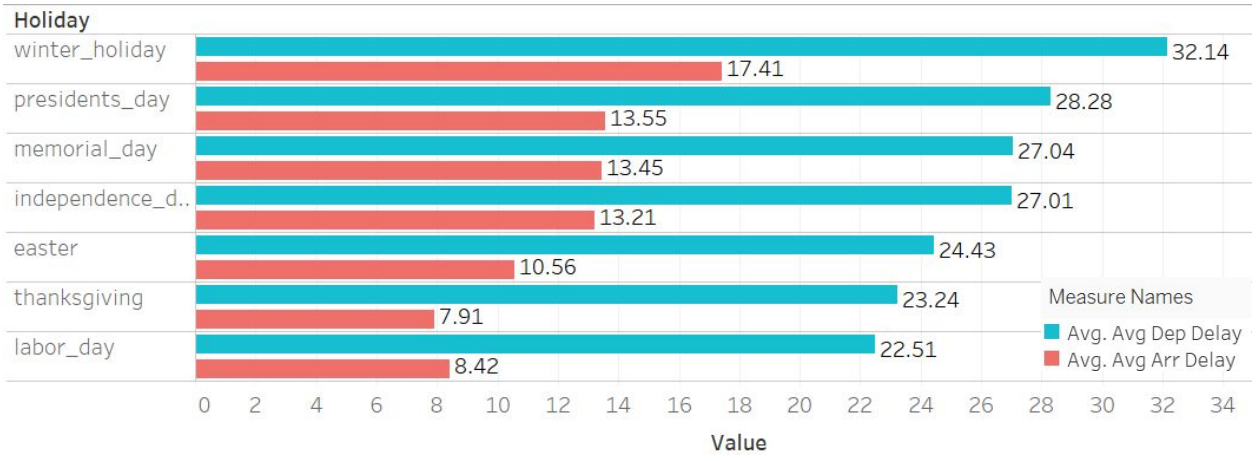
To prepare the dataset for machine learning we:

- Dropped rows containing null values. We still had a sufficient number of observations left – over 9 million records.
- Removed columns that were redundant or had similar information as well as weather variables that consisted majorly of missing values.
- Dummied the following columns: day_of_week, national holidays, month, destination airport, departure airport, weather conditions (e.g. fog), and carrier.
- Added two calculated variables: total departure delay (taxi_out delay and departure delay) and total arrival delay (taxi-in delay + arrival delay).
- Replaced values greater than the 99 percentile and less than 1 percentile by the median and the 1 percentile value respectively for our dependent variables: total departure delay and total arrival delay. This allowed us to deal with extreme values resulting from rare situations that would otherwise skew our model results.
- Due to our large size, with millions of records per year, memory limitations, and computing power limitations we decided to randomly sample 5000 rows per month per year to reduce our dataset size. By doing so we were able to increase the efficiency of our analysis while still maintaining proper representativeness of our data.
    - Departure Delay Dataset: 177,526 rows and 373 columns. This dataset includes a distance variable which is not included in the arrival delay dataset.
    - Arrival Delay Dataset: 177,526 rows and 374 columns. This dataset includes two different columns that are not present in the departure delay data set: flight speed as well as departure delay. We wanted to include the departure delay variable even though it highly correlates with the arrival delay because for the future, we intend to provide real time updates for flight status and track arrival delays using departure delay, speed, and other variables.

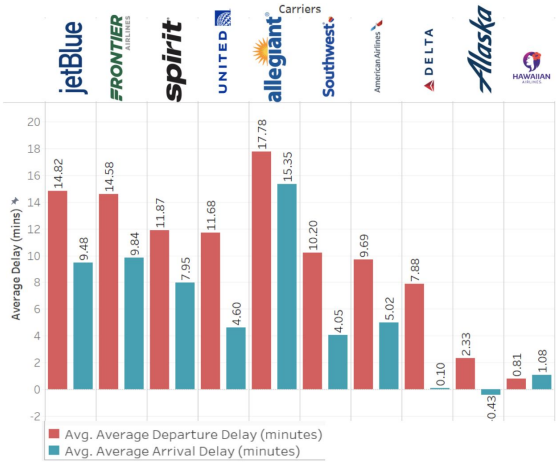
distance	distance between origin and dest airports	**thunder**	Dummy (1 or 0)	**wind_speed**	Numeric
presidents_day	Dummy (1 or 0)	**total_dep_delay**	Delay in mins	**prcp**	Numeric
easter	Dummy (1 or 0)	**total_arr_delay**	Delay in mins	**fog**	Dummy (1 or 0)
memorial_day	Dummy (1 or 0)	**dew_point_temp**	Numeric	**rain_drizzle**	Dummy (1 or 0)
independence_day	Dummy (1 or 0)	**sea_level_pressure**	Numeric	**snow_ice_pellets**	Dummy (1 or 0)
labor_day	Dummy (1 or 0)	**station_pressure**	Numeric	**hail**	Dummy (1 or 0)
thanksgiving	Dummy (1 or 0)	**visibilty**	Numeric	**month_x (12 month)**	Dummy (1 or 0)
winter_holiday	Dummy (1 or 0)	**tornado_funnel_cloud**	Dummy (1 or 0)	**Airline_x (Top 10)**	Dummy (1 or 0)
speed	Flight speed (distance/actual airborne time)	**temp**	Numeric	**Day_of_week_x (7 days)**	Dummy (1 or 0)
dest_x (for airport)	Dummy (1 or 0)	**origin_x (for airport)**	Dummy (1 or 0)		

III. Exploratory Data Analysis

- **Holiday statistics - Winter & Easter holidays have highest num. of flights delayed**

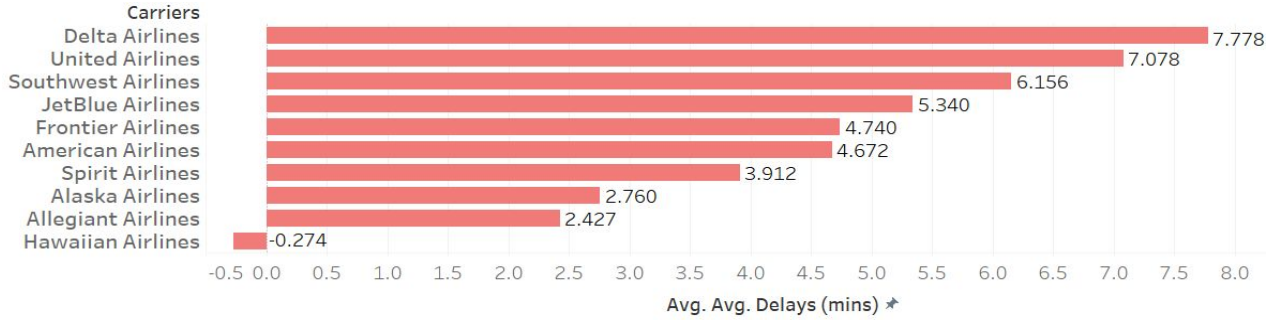


- **Flight delay summary - Airlines' Compensation of Departure Delays:**

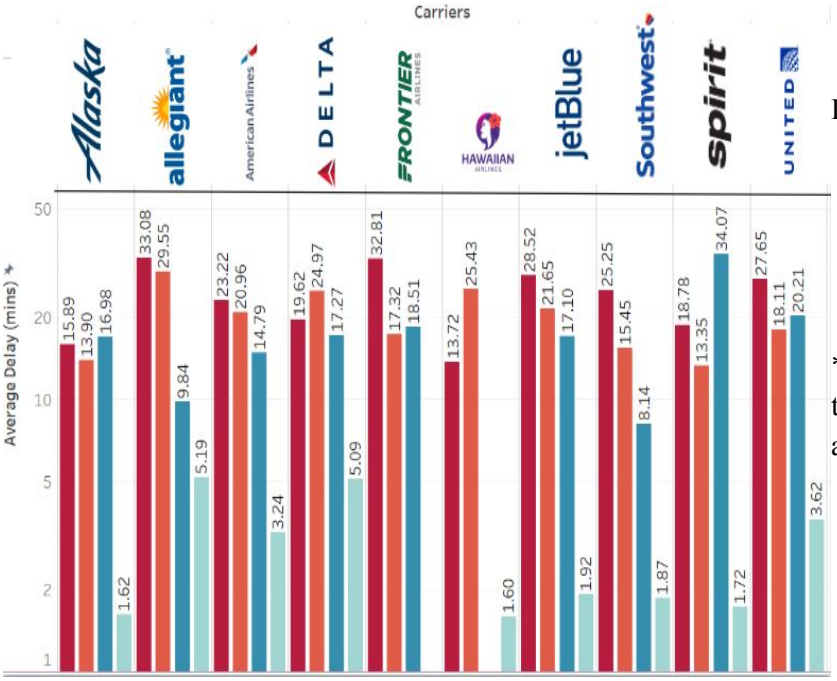


The visualization on the left represents the average delay for each airline. The red bar is the departure delay and the blue bar is the arrival delay. For the major domestic airlines, arrival delays do not reflect departure delays. Airlines with the highest departure delays are Allegiant Airlines (17.78 mins), JetBlue Airlines (14.82 mins), and Frontier Airlines (14.58 mins). Airlines with the lowest departure delays are Hawaiian Airlines, Alaska Airlines, and Delta Airlines. Their average departure delays are respectively 0.81 minutes, 2.33 minutes, and 7.88 minutes. The difference between the departure delay and arrival delay shows that arrival delays do not necessarily reflect the departure delays.

Rather, we see that typically most flights can reduce arrival delays relative to departure delays. For example, we can see that although Delta has one of the lowest departure delays it is one of the worst airlines at compensating for delays, which makes sense considering Delta's departure delays on average aren't that high to begin with. However, Allegiant is one of the airlines with the worst departure delays as well as an airline with the worst delay compensation.



- **Major Delay Reason: Late Aircraft Arrival**



Delay reasons rankings:

1. Late Aircraft Arrival
2. Carrier
3. National Aviation System
4. Weather

*security delays not included since they are typically <5 min on average

The visualization above shows the major reasons for flight delays. There are four main categories of delay reasons:

- Air Carrier Delay: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, etc.).*
- Extreme Weather Delay: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight.*
- National Aviation System Delay: Delays and cancellations attributable to the national aviation system refer to a broad set of conditions -- non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.*
- Security Delay: Delays caused by evacuation of terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and long lines in excess of 29 minutes at screening areas.*
- Late Arriving Aircraft Delay: Previous flight with same aircraft arrived late caused present flight to depart late*

**Source: Airport Traffic Bureau of Transportation*

IV. Methodology:

Using the Root Mean Square Error (RMSE) as the measure of accuracy, the team evaluated model performance on predicting future flight delays. The methodology included the following steps:

1. The team ran several flexible models including Random Forest, Boosting, and XGBoost. Cross validation was used to adjust parameters that would increase prediction accuracy. These models are slower to run but can inform the "performance ceiling" for the dataset.
2. Next, the team investigated a suite of simple, or interpretative, models: Linear Regression, Lasso, and Forward Selection. Models such as Lasso and Linear Regression are helpful in identifying the most influential variables.
3. Finally, the team identified the most influential variables by examining the most interpretable models and selected the best predictive model based on the RMSE.

V. Challenges and Solutions

- **Non-technical Challenges:**

  Due to the global pandemic, COVID-19, face-to-face meetings were no longer an option and all meetings switched to remote virtual conferences on the video communication software, Zoom. There were three main challenges encountered during the Zoom meetings:

  - First, people were easily distracted in the meetings by their surroundings and attempts at multitasking. Overall, it was difficult to maintain full attention to the meeting for a long time. We always faced the situation that people needed to catch up with the team discussions due to loss of attention. To overcome this, we only meet for 45 minutes at a time and if we meet for a longer time we take breaks in between. In addition to this, the policy of collective contribution is implemented during team discussions. Each team member is required to share his or her thoughts so that the decision will be made collectively to move on and make sure everyone is on the same page.

  - The second challenge was finding appropriate time slots that worked for everyone. To tackle this, meetings were scheduled right after the class. During the non-class week, Zoom meetings were scheduled early on to ensure everyone is able to block out the time needed for the meeting.

  - The third challenge was improving the efficiency of the team discussions. The way to guarantee productivity for each team meeting is that everyone is assigned a task to complete before the meeting. During the meeting, each team member will have a chance to show what is accomplished and then share the work with others by screen sharing. This ensured efficient use of our time and allowed us time to then brainstorm ideas and identify tasks to complete before the next meeting.

- **Technical Challenges**

  - The first challenge dealt with having to read and merge the datasets. Each annual flight data was at least GB in size. Furthermore, we needed to merge the weather dataset with the flights data. Merging the data was not possible on our local machines that had very limited computing power and memory. Therefore, the team needed to utilize cloud computing and build a VM instance with a higher CPU and memory on the Google Cloud Platform (GCP). Even with the upgraded VM instance, the R studio cloud still had issues

handling such a large dataset but provided us the necessary resources needed to merge our data and prepare for further downstream processes and analysis.

- After successfully merging the datasets, we then decided to take a fraction of the fully merged dataset, which is a multi-million-row dataset. Using only a fraction of our dataset enabled the team to perform efficient analysis on the data, given our limited computing power and memory, while still maintaining data representativeness.

- The final challenge of the project is finalizing our project objective. The team was constantly reshaping and refocusing the goal of the project. At first, our goal was to provide insights to airlines as to how they can improve their services by looking at customer complaints and evaluating an individual airline performance to other airlines in the industry. Unfortunately, the limited data we had provided little insight in identifying influential factors that can help carriers improve their services. Consequently, we decided to change our project goal in light of what data we were actually able to gather and compile.

  The first question then was "how can we deliver value to either the airline customers or airlines given the data we have?" Finally we decided to use the historical data we had and merge it with other datasets to attempt to predict future delays in hopes to provide refined delay predictions with respect to a customer' future flight.

  The second question involved dealing with the trade off between prediction accuracy and model interpretability. Methods like XGBoost might result in high prediction accuracy but low interpretability. Eventually, the team decided to use multiple models including complicated models that would allow the team to improve delay prediction's accuracy and interpretable models that would allow us to explain the variable significance and identify influential variables.

VI. Results and Discussion

- **Testing 6 models for predicting delays and examining their respective RMSE**

Model	RMSE Test (departure delay)	RMSE Test (arrival delay)
Linear Regression	26.27	14.60
Lasso Regression	25.64	14.45
Forward Selection	25.73	14.63
Boosting Trees	25.96	15.42
Random Forest	26.12	13.78
XGBoost	25.38	13.86

- **Most influential variables:**
  - Weather: (-) Visibility, (+) Snow_ice_pallette, (+) Thunder, (+) Wind_speed

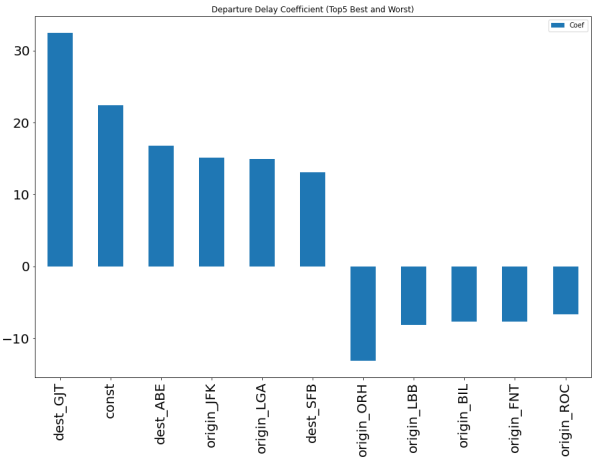
- Months: (+) May, (+) June, (+) July, (+) August
- Day of week: (+) Monday, (+) Thursday, (+) Friday
- Origin/Destination: (+) JFK, (+) SFO, (+) EWR, (+) LGA
- Airlines: (+) Frontier, (+) Jet Blue, (+) South West
- Holidays: (+) Winter Holiday, (+) Thanksgiving

Note:

(1) The influential variables are identified by combining and comparing the results of linear regression, lasso, forward selection, and random forest. Results for those models are highly consistent.
(2) + indicates a positive correlation between the variable and the flight delays
(3) - indicates a negative correlation between the variable and the flight delays

A. Linear Regression Results:

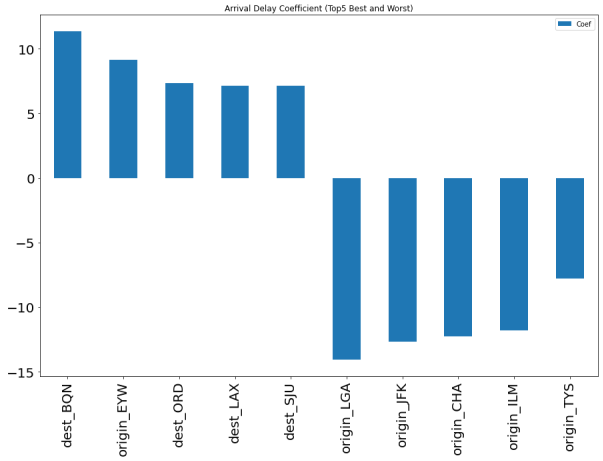
Flight Departure Delays:



For the departure delay regression model, the RMSE value is 26.27. The top 5 influential variables are origin of GJT (Grand Junction Regional Airport, 32.46 coefficient), destination of ABE (Lehigh Valley International Airport, 16.81 coefficient), origin of JFK (John F. Kennedy International Airport, 15.14 coefficient), origin of LGA (LaGuardia Airport, 14.89 coefficient) and origin of SFB (Orlando Sanford International Airport, 13.05 coefficient). That is, flights taking off from or arriving at these airports are on average having longer flight delays.

On the other hand, the top 5 negative variables are origin of ORH (Worcester Regional Airport, -13.12 coefficient), origin of LBB (Lubbock Preston Smith International Airport, -8.09 coefficient), origin of BIL (Billings Logan International Airport, -7.64 coefficient), origin of FNT (Bishop International Airport, -7.64 coefficient) and origin of ROC (Greater Rochester International Airport, -6.61 coefficient). That is, those airports have lower impact on increase of departure delay which makes sense considering they are smaller regional airports.

Flights Arrival Delays:



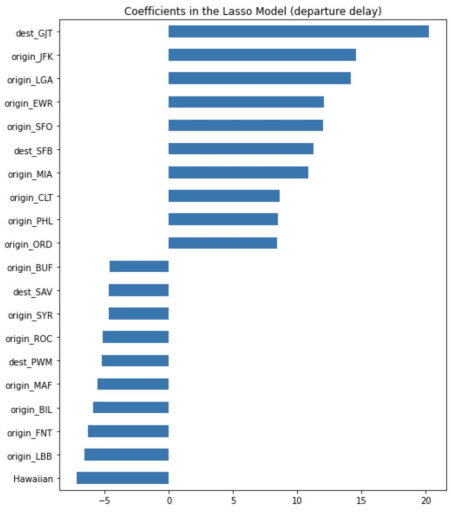
The regression model used to predict arrival delay has an RMSE value of 14.60. The top 5 positive coefficient variables are origin of BQN (Rafael Hernández International Airport, 11.34 coefficient), origin of BYW (Blakely Island Airport, 9.14 coefficient), destination of ORD (O'Hare International Airport, 7.34 coefficient), destination of LAX (Los Angeles International Airport,

7.15 coefficient) and destination of SJU (Luis Muñoz Marín International Airport, 7.13 coefficient). In other words, those airports have a greater impact on the increase of arrival delay.

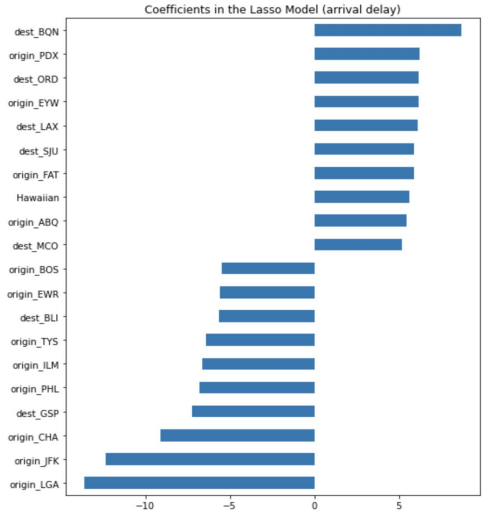
On the other hand, the top 5 negative coefficient variables are origin of LGA (LaGuardia Airport, -14.07 coefficient), origin of JFK (John F. Kennedy International Airport, -12.65 coefficient), origin of CHA (Chattanooga Airport, -12.24 coefficient), origin of ILM (Wilmington International Airport, -11.78 coefficient) and origin of TYS (McGhee Tyson Airport, -7.80 coefficient). That is, those airports have less impact on the increase of arrival delay.

B. Lasso Regression Results:

Flights Departure Delays:



Flights Arrival Delays:



The visualization above shows the top 10 positive and 10 negative significant indicators based on their coefficients. From the left-hand side, the RMSE regarding the departure delay is 25.64. The most positive coefficients are central Nebraska regional airport (GRI), John F Keenedy international airport (JFK), and LaGuardia airport (LGA). The most negative coefficients are carrier Hawaiian, Bishop International Airport (FNT), and Lubbock Preston Smith International Airport (BNT).

From the right-hand side, the RMSE under lasso regression regarding the arrival delay is 14.45. The most positive coefficients are Rafael Hernández Airport (BQN), Portland International Airport (PDX), O'Hare International Airport (ORD). The most negative coefficients are LaGuardia international Airport (LGA), John F Kennedy International Airport (JFK), and Chattanooga Metropolitan Airport (CHA).

C. Forward Selection Results:

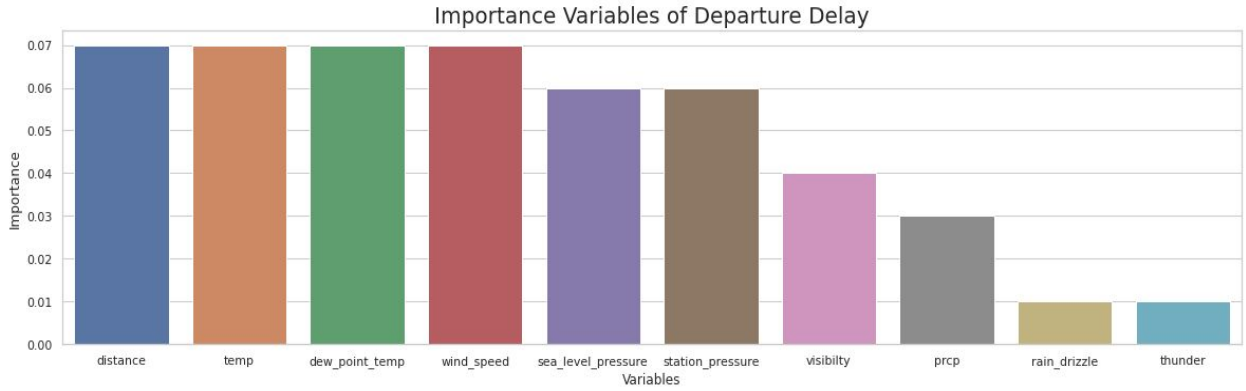
The RMSE using forward selection to predict departure delays is 26.31. Looking at the selected features before the elbow point, most influential factors include visibility, thunder, airport laguardia(origin_LGA), distance, snow ice pellets, month july (month_7), and Thursday

(day_of_week4). Those factors can be further categorized as temperature, distance, time, and airport location.

The RMSE for arrival delays is around 14, and most influential factors of arrival delay include: speed, airline Delta, origin airport Laguardia, origin airport JFK, airline United, etc.

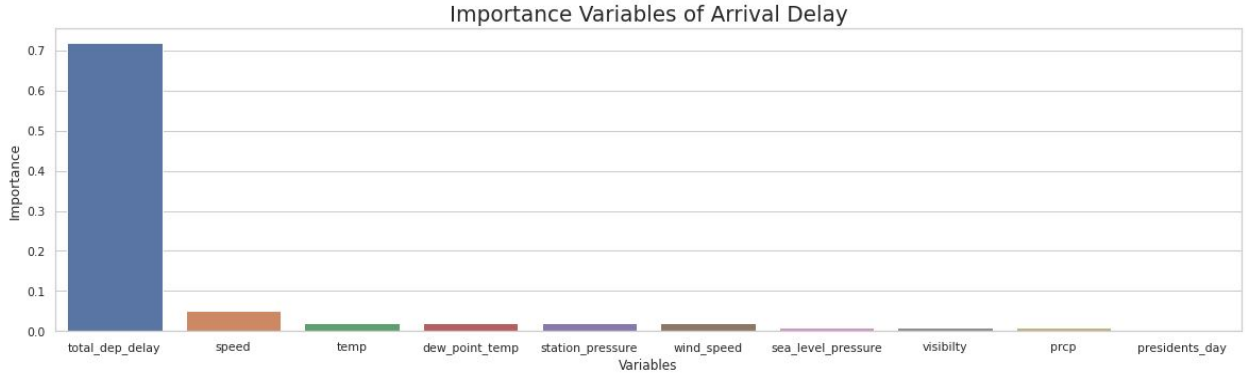
D. Random Forest Results:

Flights Departure Delays:



The Random Forest method used for predicting departure delay had an RMSE of 26.12. The number of trees used in the Random Forest model was a 100. Adding more trees beyond resulted in insignificant improvements in RMSE. The result of top 9 importance variables in the model of departure delay are distance (0.07 Importance), temperature (0.07 Importance), dew point temperature (0.07 Importance), wind speed (0.07 Importance), sea level pressure (0.06 Importance), station pressure (0.06 Importance), visibility (0.04 Importance), precipitation (0.03 Importance), rain drizzle (0.01 Importance) and thunder (0.01 Importance). In other words, those top 9 importance variables have a higher impact for the increase of departure delay in the Random Forest model.

Flights Arrival Delays:

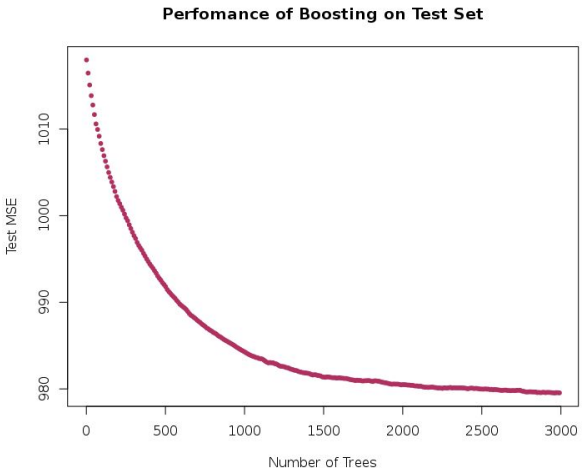


In the case of the Random Forest method, the model RMSE with arrival delay is 13.78. The number of trees in the Random Forest model that we have is 100. Adding more trees beyond

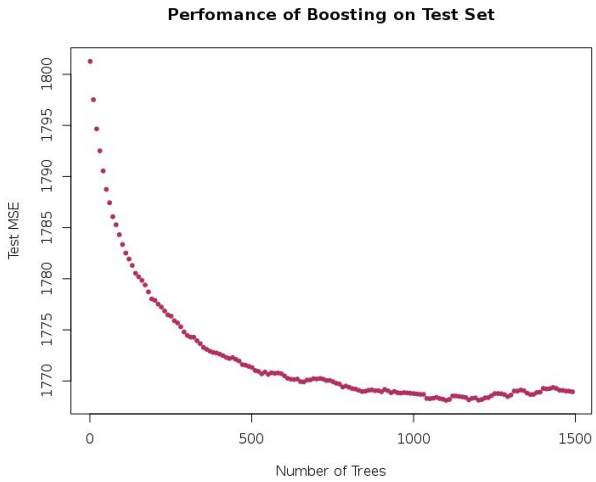
resulted in insignificant improvements in RMSE. The result of top 8 importance variables in the model of arrival delay are total departure delay (0.72 Importance), speed (0.05 Importance), temperature (0.02 Importance), dew point temperature (0.02 Importance), station pressure (0.2 Importance), wind speed (0.02 Importance), sea level pressure (0.01 Importance), visibility (0.01 Importance) and precipitation (0.01 Importance). Therefore, those top 8 importance variables have higher impact for the increase of arrival delay in the Random Forest model.

E. Boosting Results:

Flights Departure Delays:



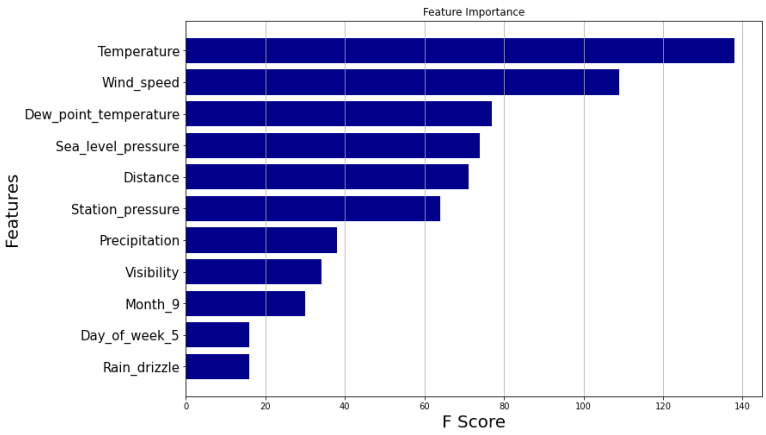
Flights Arrival Delays:



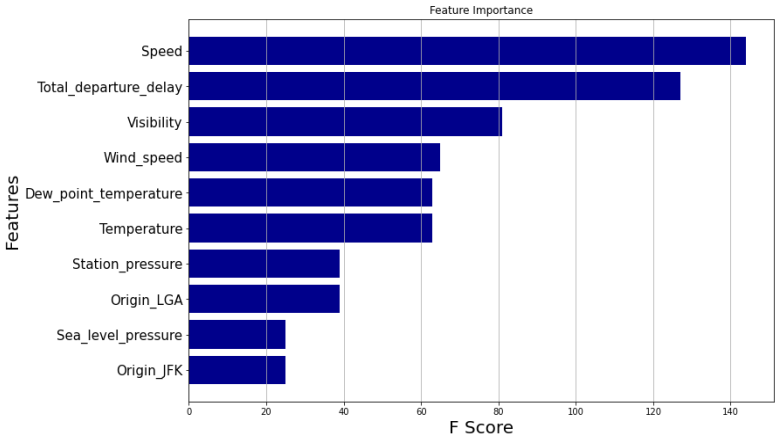
Building the boosting model for departure delays required 3000 trees while the boosting model for the arrivals delays required 1200 trees. The boosting model for departure delays resulted with an RMSE = 25.96 and the boosting model for departure delays resulted with an RMSE = 15.42. For predicting departure delays or arrival delays the top significant predictors were weather variables such as visibility and wind_speed, whether the month was June or July, and whether the airlines were JetBlue or Southwest.

F. XGBoost Results:

Flights Departure Delays:



Flights Arrival Delays:



By using a 5-fold Cross Validation model while running XGBoost predictions, we got the RMSE regarding the departure delay of 25.38 and RMSE for the arrival delay is 13.86.

For arrival delay, as we can see the feature total_departure_delay, has been given the second highest importance score among all the features, which makes sense because departure delay will absolutely cause serious influence in flight arrival delay. Other top important features regarding arrival delay, such as visibility, wind speed, dew point temperature, temperature, and sea level pressure can be considered as uncontrolled weather factors. Moreover, other important factors like the station pressure, origin of LGA (LaGuardia Airport), and origin of JFK (John F. Kennedy International Airport) are all avoidable or changeable factors while choosing flights.

For departure delay, top features are temperature, wind speed, dew point temperature, sea level pressure, precipitation, visibility, and rain drizzle are all weather issues. Other important features regarding departure delay, such as distance, station pressure, month of September, and day of week 5 are factors that customers can make changes while booking flights.

VII. Criticism of the results and future work

- First, the R square for the *Linear Regression* and the variable importance indicated by the *Random Forest* to predict departure delays are both around 7%. That is, the models can only explain about 7% of variations in our dataset.
- Second, the RMSEs for models we used were very similar to each other.
- The first two issues are probably due to multiple reasons including the facts the:
    - The majority of the variables used in our models were dummy variables, except for weather variables and the distance variable.
    - Our set of variables were limited to only weather, holiday, and route factors. As indicated by both EDA and forward selection, delays are highly correlated with late arrival (38%) and carrier (32%), and national aviation system (25%), not just unfavourable weather (5%). In reality, a lot more factors impact flight delays including airport congestion and other factors such as pandemics which are not predictable.
- To solve the problem and to improve the prediction accuracy we need to include:
    - More informative variables. For instance, Late arrival minutes - on the same date, origin, destination, and tail number - that indicate whether or not the previous flight was delayed, if so, how many minutes was it delayed.
    - Flight scheduling & airport congestion data such as delays due to flight scheduling, delays related to consumer (check-in time delays), number of flights departing and arriving at airport throughout the day etc.
    - Experiment with interactions between variables.

- Dangers of making predictions using exclusively historical data:
  The ability to predict the impending delays by studying historical data is a huge advantage. Vast amounts of historical data can help spot trends and patterns that are likely to influence future course of actions. However, the danger comes when this predictive analysis is mistaken for prescriptive analysis especially in contexts where historical data are not necessarily accurate predictors of an ever-changing world. This current model is essentially a probabilistic model, so it will not detect delays with 100% accuracy.
    - Solution: Predictive models should be constantly updated by training and validating on real-time data. Real-time predictive models have the potential to address unnecessary airport congestion and as a result, enhance the customer experience.

VIII. Conclusion

Ultimately, the most flexible model XGBoost provided the most accurate departure and arrival delays. However, to get a better understanding of the variables that are the best pfredictors for delays we need to examine a more interpretable model albeit less flexible and less accurate. By examining models such as Random Forest and Regression we can see that weather attributes followed by specific months and days as well as airlines can be the best predictors of delay.

Predicting flight delays is no easy feat. Although historical data is able to shed some insight related to the trends and seasonality in delays it definitely doesn't provide a suitable ground to build upon accurate predictions for future delays on its own. While utilizing real-time data is sure to improve the reliability and accuracy of the delay predictions, including additional variables to expand the breadth of the data is just as crucial.

References:

- https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
- https://console.cloud.google.com/marketplace/details/noaa-public/gsod?project=team5-888
- https://stat-or.unc.edu/files/2018/09/Paper3_MSOM_2012_AirlineFlightDelays.pdf
- https://www.bts.gov/topics/airlines-and-airports-0
- Data sets source: https://www.bts.gov/topics/airlines-and-airports-0
- Weather data source: https://www.ncdc.noaa.gov/cdo-web/datatools/records;
- https://console.cloud.google.com/marketplace/details/noaa-public/gsod?project=team5-888