Flight Delay Prediction using Logistic Regression

Abdullah SAHIN

Sep 12, 2020

Contents

Introduction

In the last ten years, according to the Bureau of Transportation Statistics (BTS), only 79.63% [1] of all flights have performed on time. Only a few remaining percentages were cancelled or diverted, less than 2%; rest of them were delayed mainly due to late arriving aircraft followed by the cause of the national aviation system and air carrier. A flight is considered delayed when a flight arrives or departs 15 or more minutes than the scheduled time. Averagely speaking, 720 million people [2] were on board and 144 million of those were affected by flight delays caused by five main reasons. Those reasons are [3]:

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

These series of delays cause a serious financial burden on airlines. In 2010, the Federal Aviation Administration commission estimated that flight delays cost the airline companies $8 billion a year, most of which due to increased budget on crews, fuel, and maintenance [4]. It is also worth noting that this cost adds up to $32 billion in a year by accounting for other lateral costs such as passenger cost and indirect effect on associated businesses.

By addressing delayed flights problems, it is possible to reduce its heavy financial load on airlines, passengers, and GDP. That is,

1. letting passengers know about the probability of a particular flight being delayed before booking can help them to plan on a better date and time. In addition, it will save passengers more than a million hours of unnecessary delays per year [5].

2. booking websites, airline companies, travel agencies can provide better customer service. Providing likelihood of flight delays to clients can decrease the number of complains and increase customer satisfaction.

3. airline companies can pinpoint important factors that cause flight delays and take precautions depending on this information. The principal benefit of this study to airlines is the reduction in additional operating cost caused by delays.

Overall, the findings of this study can provide a high-profile achievement by addressing aviation delay problems with a robust prediction model and helping people and businesses better on planning their flights.

Data Acquiring & Wrangling

In this section, we will provide information about what data acquiring and data wrangling processes are taken.

Data Collection: For this study, we used four different datasets from various sources.

1) Airline on-time performance dataset is available on The Bureau of Transportation Statistics' website. In the process of data scraping, I used the selenium library in Python. This library works with a specific version of a Chrome driver that needs to be compatible with your chrome version and can be downloaded from this page. It downloads the monthly data to the ~/data/flight_data folder. From then on, a shell script concatenates these monthly flight data into a single csv file under the same folder.

2) Weather data was obtained from Iowa State University's Environmental Mesonet Platform. This platform works like an API service which needs modification of a requested URL for each inquiry. The python script for weather data can be found on this GitHub repo.

3) Airport data was obtained from The Bureau of Transportation Statistics' website manually. This dataset has a substantial amount of information such as airport names, location, time zone etc.

4) ICAO data was acquired from OpenFlight dataset. The reason why we need this data is because The Environmental Mesonet Platform of Iowa State University requires ICAO codes as station IDs whereas airport data from The Bureau of Transportation Statistics have IATA his information does not exist in the airport dataset obtained from BTS.

Data Cleaning: Airline on-time performance dataset have redundant out-of-scope information, so I only kept 30 necessary features in the data gathering process. I created a dictionary called flag to identify null entries (condition 1), out of range entries (condition 3), and both null and out of range entries (condition 2). Certain implementations have been performed based on flag values. For instance, a weekday column has to be in between 1 and 7. Any entry that has null in that column needs to be imputed using pandas dt accessor. Here is the list of ranges of certain columns.

Column	**Minimum**	**Maximum**	**Treatment (or Remarks)**
Date	Given	Given	Any date entries out of a given range is removed.
Weekday	1	7	Imputed if possible (using dt accessor of pandas)
IATA	N/A	N/A	Any null entries are removed.
TailNum	N/A	N/A	Any null entries are removed.
OrgAirID	10001	16878	Any out of range entries are removed.
DestAirID	10001	16878	Any out of range entries are removed.
OrgMarID	30001	36845	Imputed if possible. Otherwise removed.
DestMarID	30001	36845	Imputed if possible. Otherwise removed.
Div	0	1	Any wrong entries are imputed based on certain columns.
Cncl	0	1	Any wrong entries are imputed based on certain columns.
CnclCd	1	4	Any out of range entries are removed.
ScDepTime	0	2400	Any out of range entries are removed.
ScArrTime	0	2400	Any out of range entries are removed.
ScElaTime	N/A	N/A	Any null entries are removed.
DepTime	0	2400	Imputed if certain conditions met.
DepDelay	N/A	N/A	Imputed if certain conditions met.
TxO	N/A	N/A	Imputed if certain conditions met.
TxI	N/A	N/A	Imputed if certain conditions met.
WhOff	0	2400	Imputed if certain conditions met.
WhOn	0	2400	Imputed if certain conditions met.
ArrTime	0	2400	Imputed if certain conditions met.
ArrDelay	N/A	N/A	Imputed if certain conditions met.

Table 1: Data Imputation

In the final form of the data, null entries are only allowed when a flight is cancelled or diverted. The columns need to be null when a flight is cancelled are:

DepTime, DepDelay, TxO, WhOff, WhOn, TxI, ArrTime, ArrDelay, AcElaTime, AirTime

and when a flight is diverted are:

ArrTime, ArrDelay, AcElaTime, AirTime, TxI, WhOn

From the following figure below, one can understand the flowchart of data wrangling process.

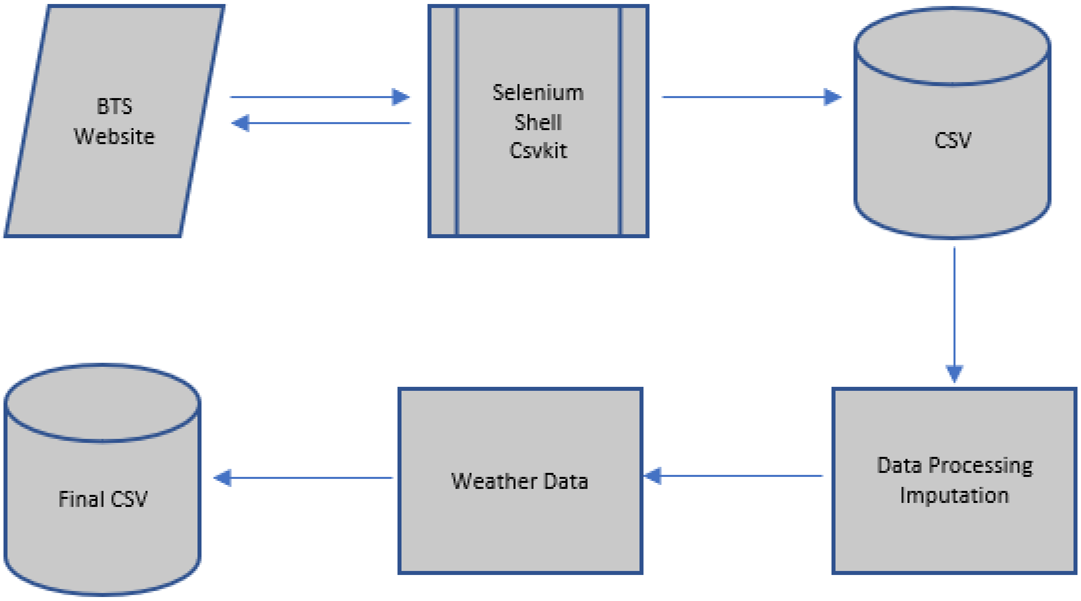

Figure 1: Flowchart of data wrangling processes

Other than, certain situations listed above, the dataset is forced to be not null and consistent. Data integrity is also checked under data wrangling Jupyter Notebook. In this part, we checked the data integrity mainly for timestamp and timedelta data types in the dataset.

Exploratory Data Analysis

In this part, data importing is performed along with down casting of columns to save some space in memory. We are going to use the flight data of year of 2018. All other miscellaneous data can be found under ~/data/misc folder.

Busiest Airlines and Airports

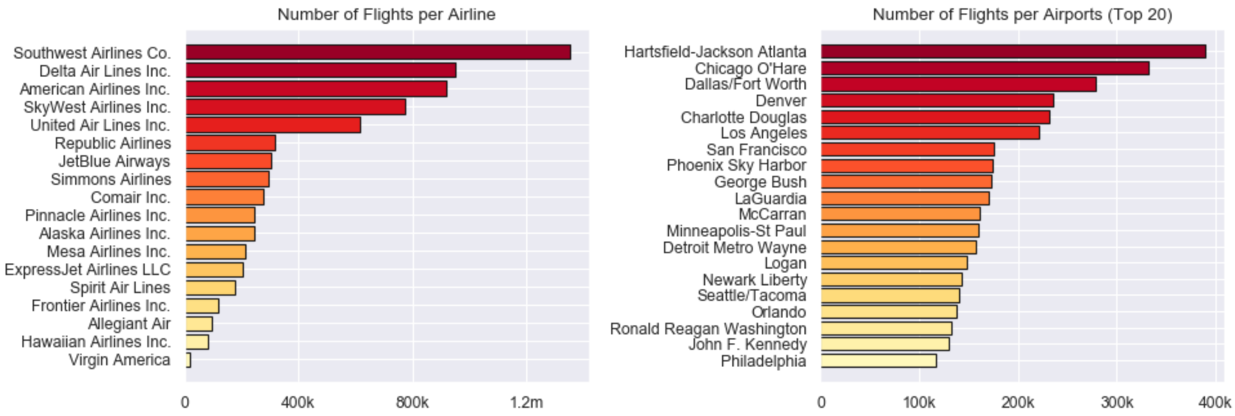

Figure 2: Busiest Airlines and Airports

Figure 2 show the most flying 20 airlines in the US on the left and the busiest 20 US airports on the right. With more than 1.2 million flights during 2018, Southwest Airlines is the most flying airline that is followed by Delta and American Airlines with 900k+ flights. Furthermore, Virgin America is the least flying airline with less than 50k flights. On the other hand, Hartsfield-Jackson Atlanta Airport is the busiest one with almost 400k flights during 2018 that is followed by Chicago O'Hare with 330k flights. They are the only two airports passing 300k flights boundary. There are only 4 airports (Dallas, Denver, Charlotte, and Los Angeles) that have 200k-300k flights in a year.

Airline On-Time Performance (General)

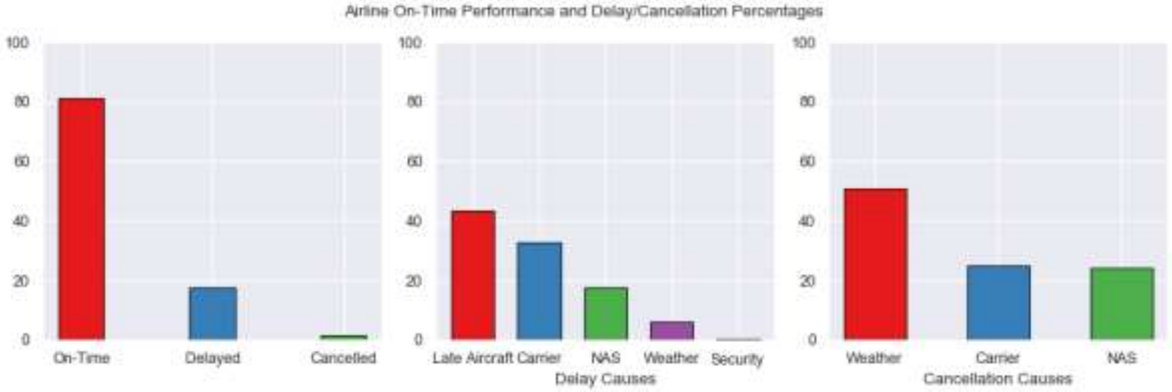

Figure 3: Airline On-Time Performance (General)

There are several causes for a flight to be delayed. It is interesting to look at the proportions or chances of encountering certain delay type. Almost 20% of the fights tend to be late (this taking account delays more than 15 mins). Most planes are late due to late aircraft and/or airline delays. This is interesting because it means that it is an area of the airline industry that could be optimized. Supporting our argument, one notices that the lower amount of delays is caused due to weather and security issues. Showing us that security for example is optimized to a very large level and other types of delays could eventually be avoided by optimizing organization in airports or establishing a delay prevention schedule. Weather delays are rather rare and are not primary reasons for flights to be delayed.

Airline On-Time Performance per Airline Companies and Airports

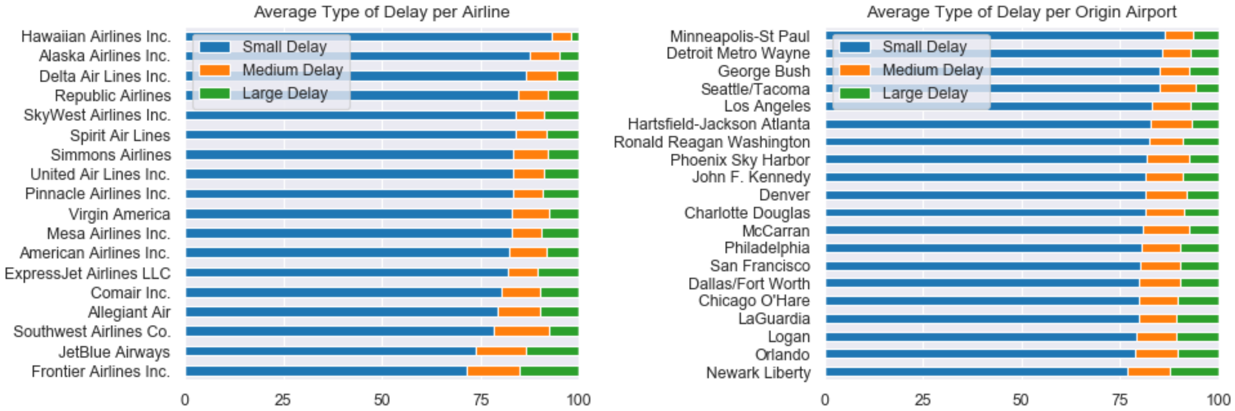

Figure 4: Airline On-Time Performance per Airline Companies and Airports

Delays are divided into three categories, namely "On Time or Small delay" (up to 15 minutes delay), "Medium delay" (15 – 45 minutes delay) and "Large delay" (45 minutes delay). In this way the graphic representation is more understandable as well as the possibility of directly comparing the variables related with delays. As represented, one can observe that Airlines that were particularly good at being on time compared to other airlines were Hawaiian Airlines, Alaska Airlines, and Delta Airlines. Those that tend to have a larger delay are Frontier Airlines, JetBlue Airways, and Southwest Airlines.

The same type of analysis can be done for different airports. In this case one can observe that airports such as Minneapolis and Detroit stand out positively, while the one with the worst frequency of delays is Newark. However, the differences are less obvious, and one can conclude that airlines play a more significant role than the airports of departure.

Since, all the airlines are operating in these top 20 airports, the delay variation due to a specific airline is averaged out. Also, the busiest airports shown in figure above are not the best performers due to the domino effect of a delay caused by environmental conditions.

Effect of departure time and weekdays of a flight on departure delay

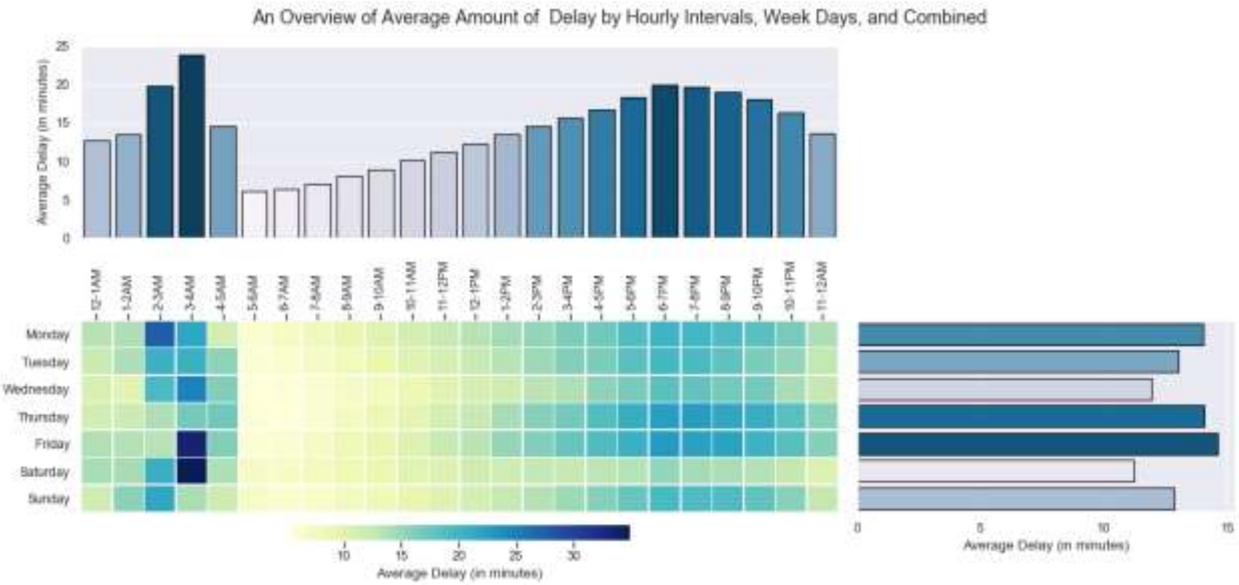

Figure 5: Effect of departure time and weekdays of a flight on departure delay

Figure 5 summarizes average amount of delay in minutes by daily, hourly, and combined time intervals. One can realize that the two histograms around the heat map represent average delay with respect to daily (on the right) and hourly (on the top) axes. It is clearly seen that 3:00AM-4:00AM time slot is the worst time to fly, especially on Friday and Saturday nights, due to long average delays (>30mins). On the other hand, it is best to schedule a flight in the early morning (5:00AM - 10:00AM), because the average delays are way below the yearly average (10 mins). Towards the evening, the delays are monotonously increasing and reaching its peak at 6:00PM-7:00PM. One can interpret this graph as the sum of two gaussian distributions centered at the two peak delay times mentioned above. The standard deviation during the night is much smaller than the evening. I believe this is related to the low density of flights during the night.

Day to day variation of the delays is much smaller (< 4mins) than the variation during the day. The best days to travel are Saturday and Wednesday that can be associated with weekend-oriented travels where Friday evening and Monday morning are busier. Due to seat availability and cost, the increase in mobility is extending to Thursday and Tuesday. Long weekends may also have some effect in this increase.

Trends by Quarters

The graph below shows average amount of delay per day in four quarters of a year. Blue shaded area represents plus and minus five days from national holidays (in total 10 days). Federal Aviation Administration (FAA) considers any flight that is late more than 15 minutes as a delay. That is why I also emphasized 15 minutes with a dashed line in the plot. The increase in average delays before and after the national holidays can be associated with higher traffic density of the airports. This can be seen in the following Figure 7.

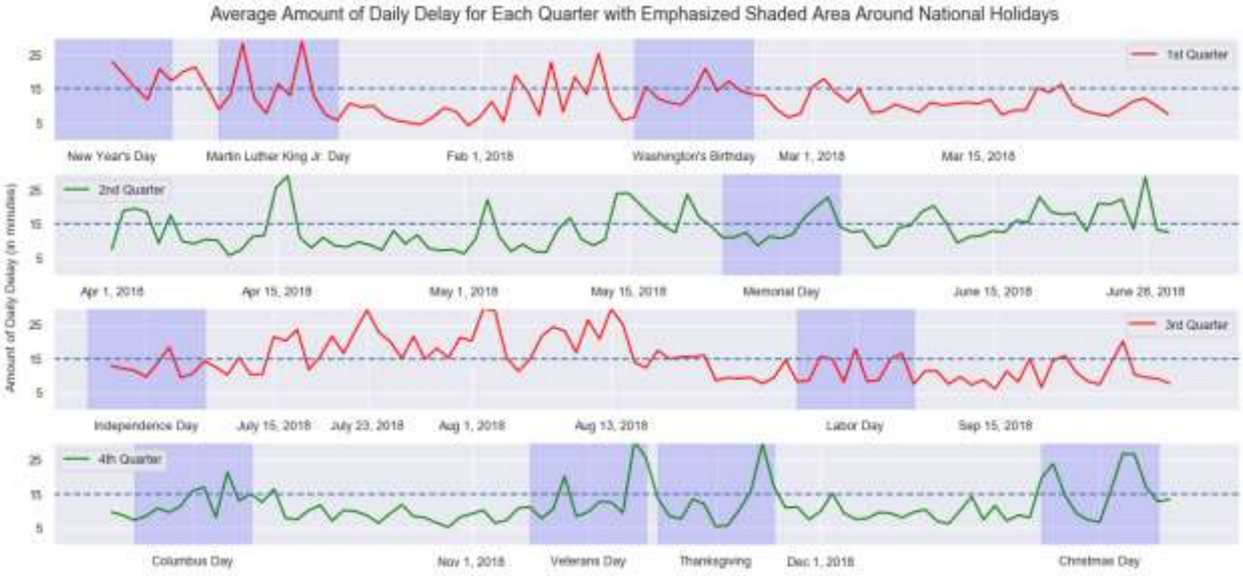

Figure 6: Trends by Quarters

Effect of Total Number of Flights on Average Amount of Delay

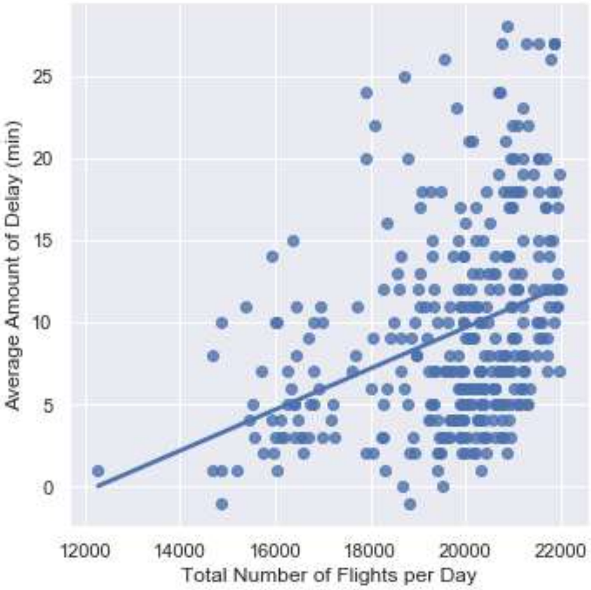

Figure 7: Effect of Total Number of Flights on Average Amount of Delay

Figure 7 shows the trend analysis that investigates the effect of airport traffic density on average delay. The solid blue line shows the trend while dots represent the individual days of year 2018. It is clearly seen that there is a linearly increasing trend as expected.

National Holiday's Effect on Total Number Flights

The number of flights during 2018 are plotted daily to understand is there any density increase during the national holidays. It seems that during the national holidays the flight distribution is redefined, but there is no visible increase in the number of flights. During normal times, there is a pattern that repeats itself weekly with a valley on Saturdays. However, the pattern is changed during national holidays according to the day of the holiday. The most drastic change happened during Thanksgiving holiday where Thanksgiving Day (Thursday) has the lowest number of flights of the entire year and Saturday of that week has way more flights than average number of flights on Saturdays. Additionally, the number of flights is varying with respect to the seasons. During autumn and winter, they are generally low and increases in Spring and reaches maximum point in mid-summer. This is probably related to summer break of students and vacation preference of workers.

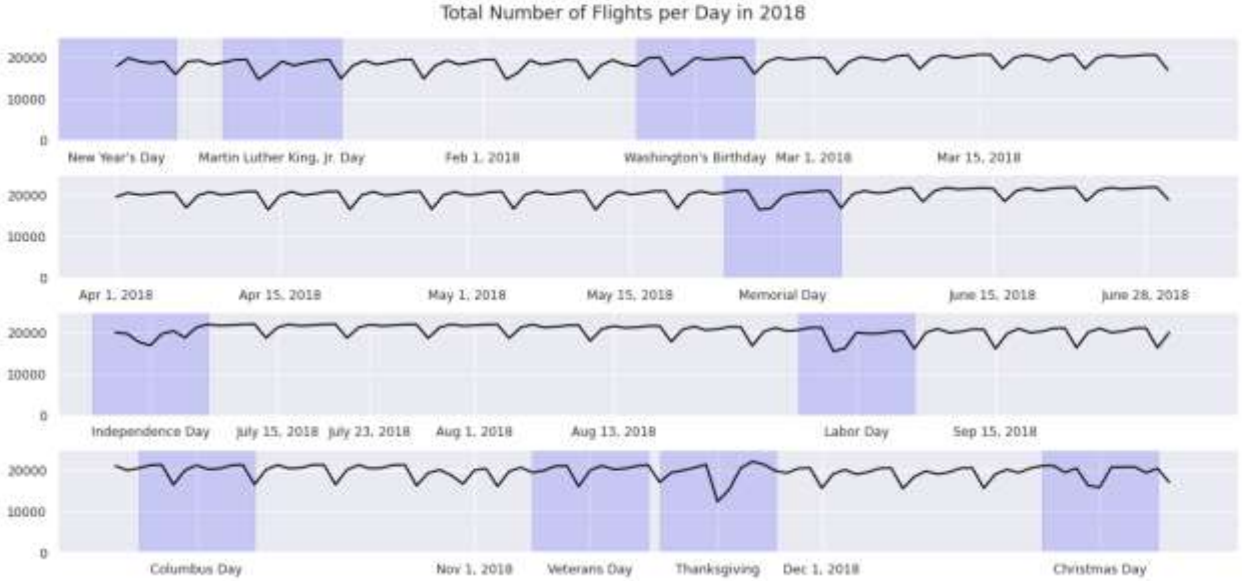

Figure 8: National Holiday's Effect on Total Number Flights

Total Number of Flights (Daily) Distribution per Weekday

Figure 9 shows the density of the flights for each day of the week for entire year. Therefore, each day has 52 data points demonstrated as both a boxplot and a violin plot that is symmetrically plotted Gaussian distribution of points. It is clearly seen that Saturday is the least dense day of the week. Also, the low end of the Thursday and Friday are related to the Thanksgiving week. Additionally, low end of Sunday and high end of Monday are probably related to long weekends where some travel is shifted to the next day.

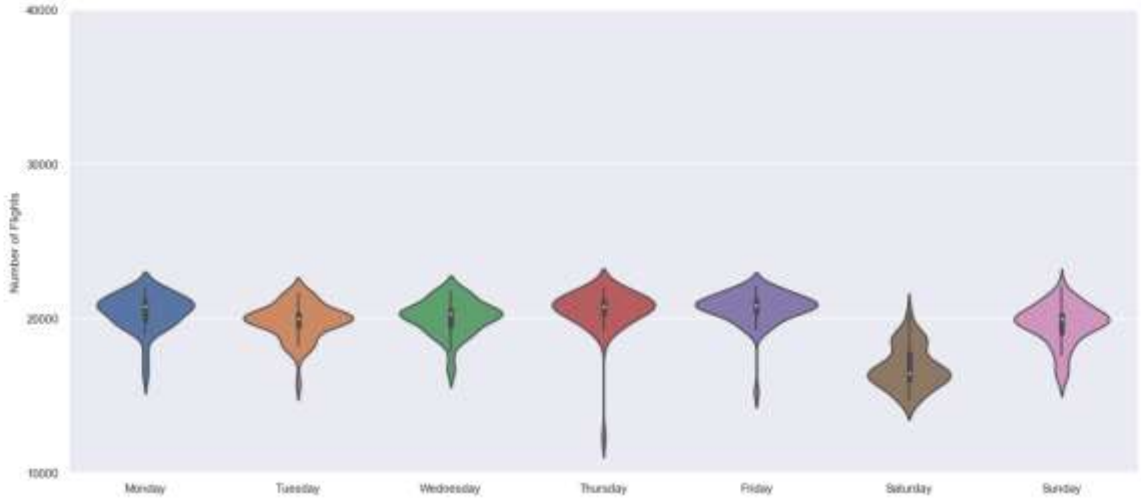

Figure 9: Total Number of Flights (Daily) Distribution per Weekday

Effect of Weather Conditions on Delays and Cancellations

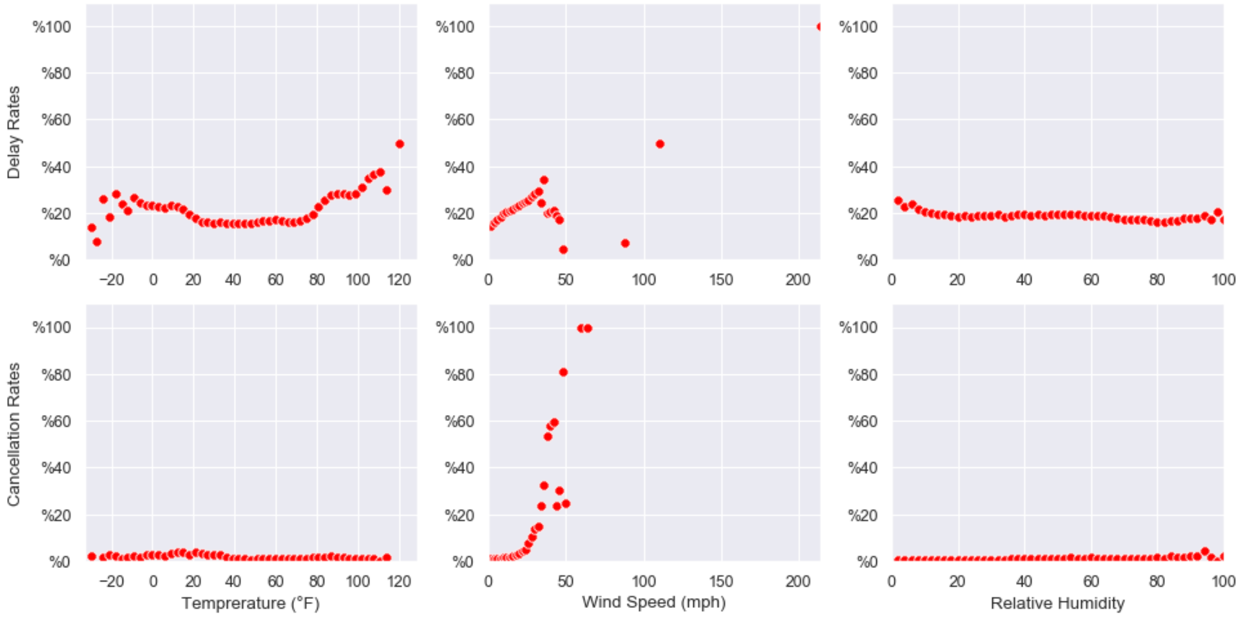

Figure 10: Effect of Weather Conditions on Delays and Cancellations

In this Figure 10, the effect of different weather parameters such as temperature, wind speed and humidity on delays and cancellations is demonstrated. It seems that the delays and cancellations are almost indifferent for increase in temperature and humidity. High temperatures above 80 causes more delays than base delay of 15-20% that is probably due to the other delays unrelated to weather. On the other hand, delays are not monotonously increasing in low temperatures. The variation is probably due to the experience and presence of the equipment to land and take off planes in the airports where freezing weather is expected.

In is clearly seen that wind speed is directly influencing the delays and cancellations. The delays are increasing linearly up to 40mph. Then, the delays are behaving different. Main reason is the exponential increase of cancellations with the wind speed. Since most of the flights above 40mph are cancelled, the delay rate is decreased. Also, there are delays at higher speeds where no cancellation data is given. They are probably due to the short duration of such high-speed winds.

Effect of Wind Speed and Direction on Weather Caused Delays

In this Figure 11, the maximum wind speeds for each direction observed in top 12 US airports are shown as a black solid line. The colored dots represent the amount of weather caused delay for corresponding wind speed and direction measured. In this figure, green dominant areas show that airport operation continues well under corresponding weather. Therefore, airports like Chicago O'Hare, Hartsfield-Jackson Atlanta, Dallas Fort Worth and George Bush are operating well for most of the wind speeds. On the other hand, Los Angeles, San Francisco, Phoenix Sky Harbor and McCarran are affected by even relatively low speeds. This difference is related to the design of the airport including its position, direction and length of the airport runways, type of the winds, and technological advancement of the airport. For example, Los Angeles and San Francisco are placed near ocean, therefore the winds from ocean can cause more problem even at low speeds due to immediate weather change. On the other hand, Chicago O'Hare is a large airport with several runways and still expanding. Its history shows that runway lengths and directions are changed to operate at high speed winds from North.

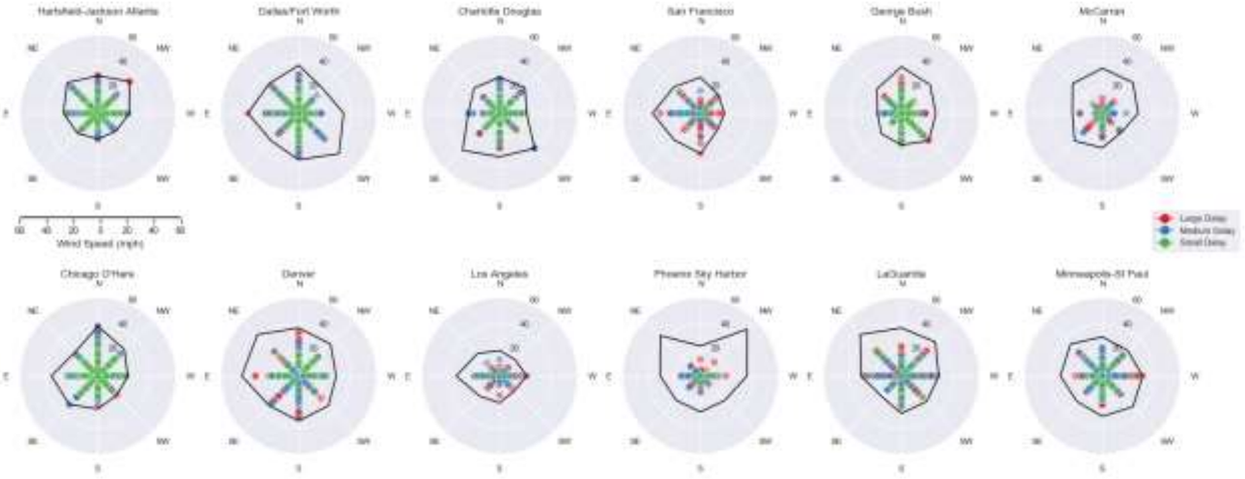

Figure 11: Effect of Wind Speed and Direction on Weather Caused Delays

Effect of Wind Speed and Direction on Weather Caused Cancellations

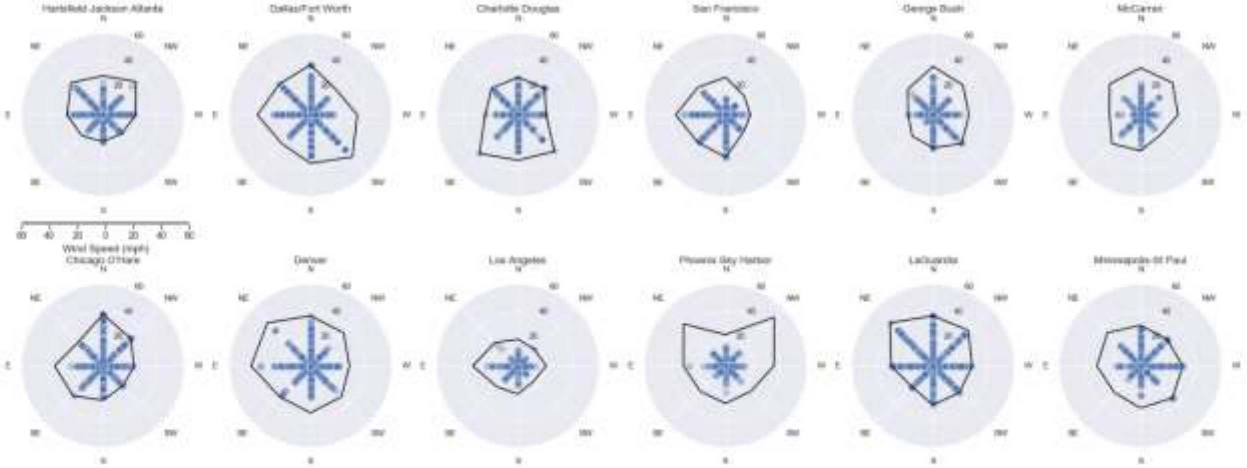

Figure 12: Effect of Wind Speed and Direction on Weather Caused Cancellations

The Figure 12 shows the weather-related cancellations similar to Figure 11. There are cancellations for every wind speed and direction for each airport. The distribution is similar to the delays with extra points at high wind speeds. As an example, all flights in LaGuardia airport are cancelled when the wind from North East direction exceeded 30mph. It is not easy to comment on the cancellations at low speeds where normally small delays are expected. Because the cancellation rate is approximately 1-2% as shown in Figure 10.

Origin-Destination Pair Average Amount of Delay

Average amount of delay of the flights operated between top 20 airports are shown as a heat map where rows are labeled according to the origin airport. Grey cells show NaN values meaning no flight data is available between corresponding airports during 2018. It is obvious that Newark Liberty is the worst destination airport. Additionally, the flights between Chicago O'Hare and John F. Kennedy airports have larger delays than the average in both directions. With using heat map, one can compare on time performance of nearby airports and, if possible, select destination and origin airports accordingly.

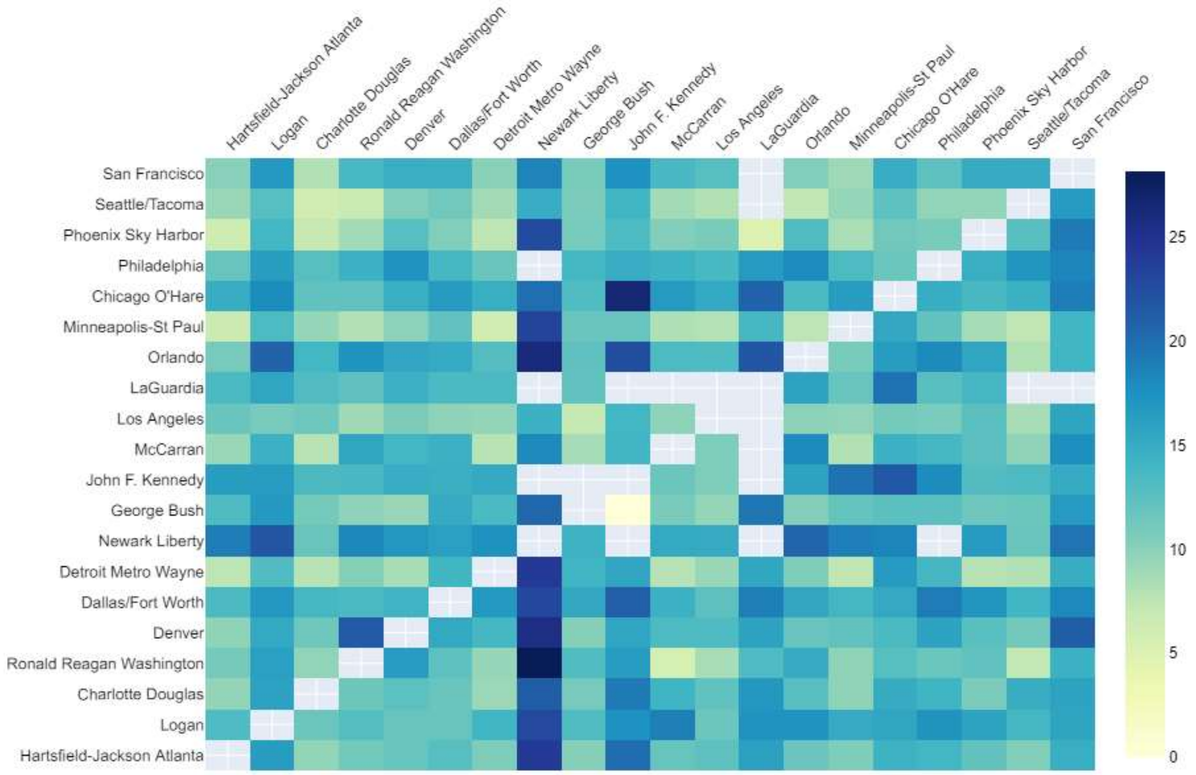

Figure 13: Origin-Destination Pair Average Amount of Delay

Late Aircraft Delays

The graph on the left demonstrates how the delay on the arrival affects the departure of the next flight of the same aircraft. The reason of the departure delay is color coded. On time departures, shown as orange, occur even when there are large arrival delays. It is probably related to scheduling where the plane has several hours before next flight. Furthermore, Late Aircraft Delay (shown as orange) influences the next flight in a linear fashion as expected. Moreover, weather and carrier related delays, shown as green and red respectively, are the main reasons of the delays when the aircraft is completed previous flight on time.

On the other hand, the right figure shows the effect of departure delay from the origin airport on the arrival of the same flight to the destination airport. The data behaves linear as expected, because the travel speeds and routes of flights are well regulated. Still they can speed up or take different routes with the permission of aviation authorities to go faster than planned. In contrast, they may need to slow down, take longer path due to weather or wait some time to land due to heavy airport traffic that will cause extra delays. The standard deviation of the delays from linear line is 13 minutes.

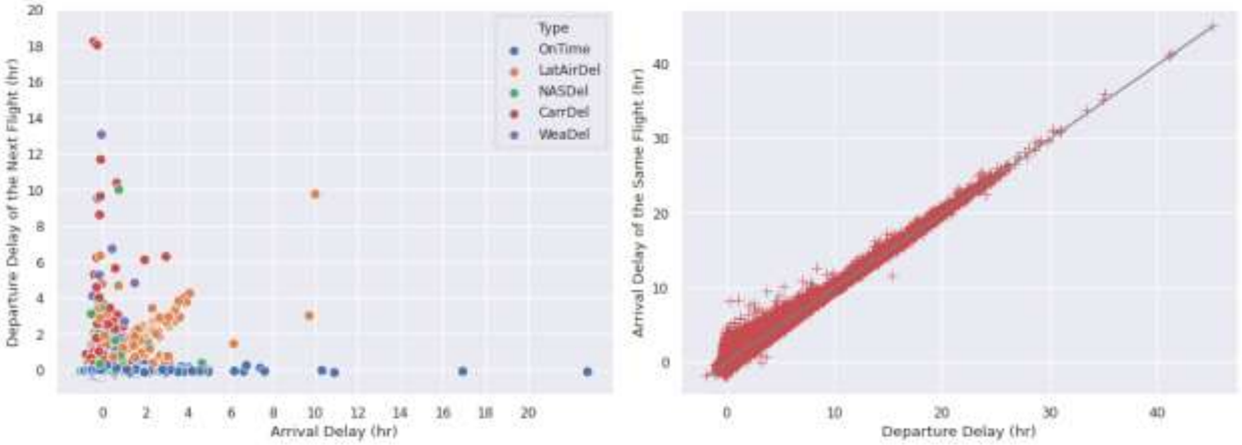

Figure 14: Late Aircraft Delays

Data Preprocessing and Feature Engineering

In this segment, we will extract various features to construct our model for anticipating the takeoff delays on flights. We will utilize those features from our dataset that contribute to the forecast. For some of the categorical data, we chose to use one hot encoding method while few of them remained as it is. On the other hand, we scaled continuous data between 0 and 1 to speed up convergence in logistic regression. We mostly referenced our Exploratory Data Analysis when extracting features and used less intuition.

Time: We start with first and foremost feature for a delayed flight: time-related data. In this category, extracted month, weekday, day of the month, day of the year, time of the day. These features will be dummy coded in the modeling part. One can see importance of these features from Figure 5 and Figure 6.

Holidays: At first, we thought that the more you are close to certain holidays, the more likely flights are delayed. To show this point, we generated Figure 6 and shaded plus/minus five days around federal holidays. For particular holidays, such as New Year's Day and Thanksgiving Day, average amount of delay is increased set threshold, 15 minutes. We decided to use these features as it is without using dummy variable.

Number of Flights: After carefully evaluating Figure 6, we decided that holidays might not be the actual cause for a flight being delayed. The actual reason could be the number of flights increased around holidays. Figure 8 demonstrates that number of flights is not a function of how close a flight to the federal holidays. However, from Figure 7, one can say that number of flights can cause average amount of delay. We extracted number of flights during time of the day for each airport.

Airline and Airports: Figure 2, Figure 4, and Figure 13 show that certain airlines and/or airports have better performance than others. IATA codes and airports codes will be one hot encoded for this feature.

Weather: Extreme weather conditions have an adverse effect on flight delays and cancellation. Figure 10, Figure 11, and Figure 12 can support this point of view. Therefore, we will be using temperature, dew point, relative humidity, heat index, wind direction, wind speed, and visibility columns.

Late Aircraft Delay: Almost half of the flight delays caused by late aircraft, Figure 2. It is crucial to take into account this effect. Also, Figure 14 show that there is linear relationship between arrival delay and departure delay of next flight (orange dots).

We listed all the feature we used in logistic regression and explained each of them in the table below.

Feature Name	Type	Usage	Explanation
month	Categorical	OHE	Month of the year
weekDay	Categorical	OHE	Weekday
dayoftheMonth	Categorical	OHE	Day of the month
dayofYear	Categorical	OHE	Day of the year
hour	Categorical	OHE	Hour of the day
iata	Categorical	OHE	IATA code of an airline
orgAirport	Categorical	OHE	Origin airport code
destAirport	Categorical	OHE	Destination airport code
temp	Continuous	As it is	Temperature
dewPoint	Continuous	As it is	Dewpoint
relHum	Continuous	As it is	Relative humidity
heatInd	Continuous	As it is	Heat index
windDir	Continuous	As it is	Wind direction
windSp	Continuous	As it is	Wind speed
visib	Continuous	As it is	Visibility
n_flights	Continuous	As it is	Number of flights per hour of the day
NewYearsDay	Continuous	As it is	Days away from New Year’s
MartinLutherKingJrDay	Continuous	As it is	Days away from MLK Day
WashingtonsBirthday	Continuous	As it is	Days away from Washington Birthday
MemorialDay	Continuous	As it is	Days away from Memorial Day
IndependenceDay	Continuous	As it is	Days away from Independence Day
LaborDay	Continuous	As it is	Days away from Labor Day
ColumbusDay	Continuous	As it is	Days away from Columbus Day
VeteransDay	Continuous	As it is	Days away from Veterans Day
Thanksgiving	Continuous	As it is	Days away from Thanksgiving
ChristmasDay	Continuous	As it is	Days away from Christmas Day
NxtNewYearsDay	Continuous	As it is	Days away from next New Year’s
prevArrDel	Continuous	As it is	Arrival delay of previous flight
timeDiff	Continuous	As it is	Time difference between arrival time and departure time

Table 2: Model Features

Modeling

Before running machine learning algorithms, it is wise to slice our problem. This can be done either per origin-destination airport pair or just per airport basis. There are also several other options can be used depend on what we want to predict. Slicing can reduce the number of the features along with better prediction. We chose to slice our problem per origin airport and month of the year basis. After grouping data per origin airport and month of the year, there are few things need to be done on the data. We outlined those steps below:

One Hot Encoding: Since we slice our problem per origin airport and month of the year, we do not need month column and origin airport column. We applied pandas get_dummies method on the remaining categorical variables.

Data Splitting: The second step involves splitting the label encoded dataset into train and test datasets. In this project we split them equally with 75%-25% ratio. Also, we split them in such a manner that the fractions of both classes remain almost same in train and test datasets.

Resampling: Since out data is imbalanced, where classes in the target variables is not distributed equally, we try random over sampling techniques to overcome this issue. One can also use random under sampling technique or other techniques such as SMOTE or SMOTENN in imbalanced-learn library, we decided to keep computational expenses low at this stage.

Hyperparameter Tuning with Cross Validation: Sometimes, rather than resampling, weighting the training samples works best using grid search algorithm. In the weighting technique, more weights are given to minority class examples. In our preliminary results, we see that random over resampling and hyperparameter tuning work best for logistic regression, so we stick with this method among all models.

Scaling: Instead of using standard scaler from sklearn framework, we used maximum absolute scaler to keep our dummy variables same.

Logistic Regression (Baseline Model)

We start with baseline model with logistic regression. In this model, we did not try hyperparameter tuning with cross validation and resampling preprocessing methods. The way how we slice our problem is per top 20 airports and 12 months of the year. Therefore, we have 240 different logistic regression model for each subset of the dataset. A convenient way to compare our result of the baseline model with others would be using their histogram of performance metrics and their mean values. Let us start with looking at their means.

Logistic Regression (Baseline)	Average Performance Metrics
Precision (Delayed Class)	0.80
Recall	0.50
F1-Score	0.61
AUC	0.83
AP	0.76
Training Accuracy	0.79
Testing Accuracy	0.79

Table 3: Results of Baseline Model

Results table provided above shows average metrics of all 240 different models. Even though training accuracy and testing accuracy seem that the baseline model performed well, the majority class will overwhelm the number of examples in the minority class, meaning that even unskillful models can achieve higher accuracy scores, depending on how severe the class imbalance happens to be. Therefore, we need to focus on other metrics such as precision and recall. Precision quantifies the number of positive class predictions that belong to the positive class whereas recall quantifies the number of positive class predictions made of all positive examples in the dataset.

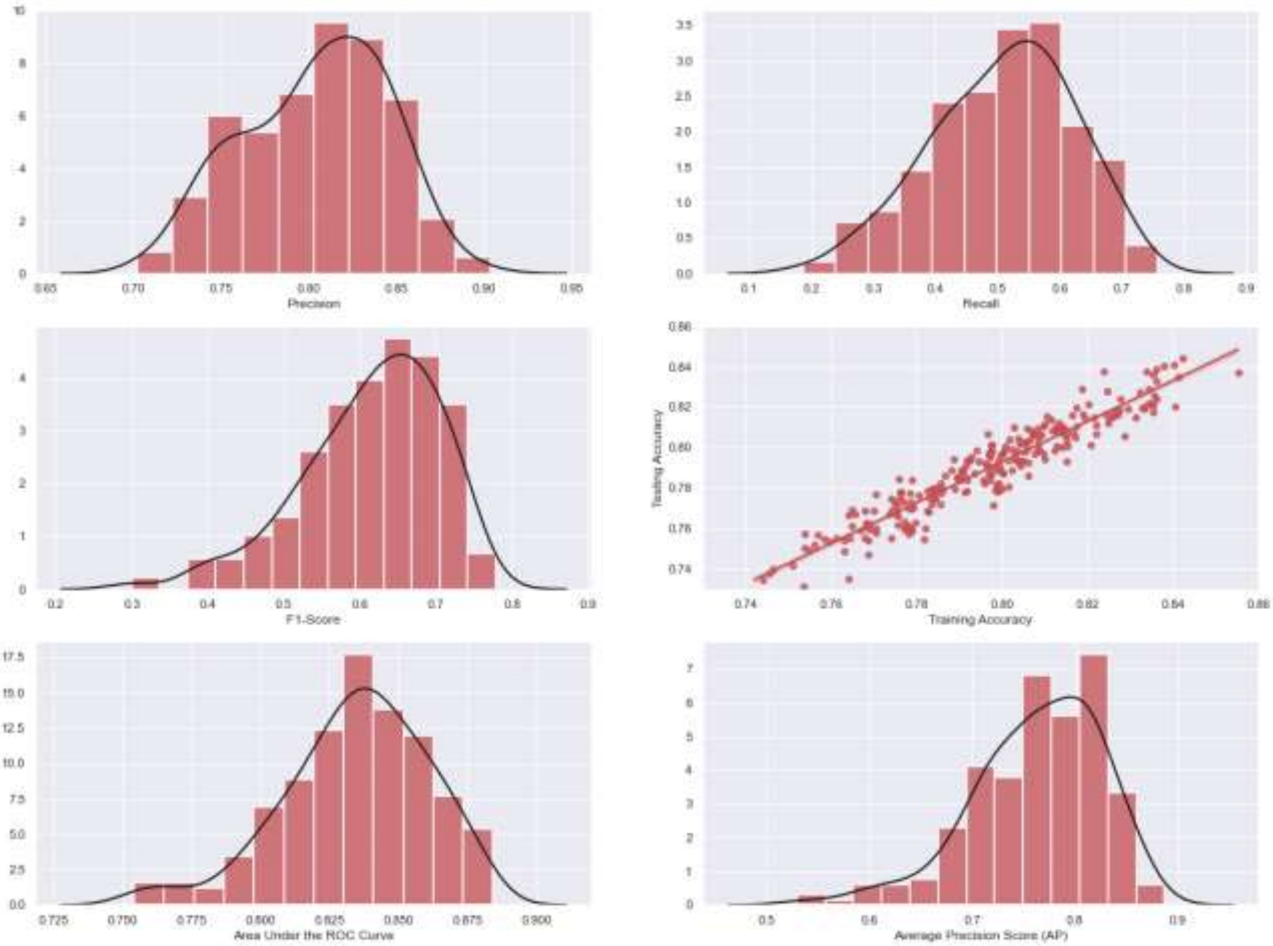

Figure 15: Histogram of Results (Baseline Model)

From Figure 15, one can see that precision scores are high whereas recall scores remains low. This can be explained by our baseline model behaves meticulous and does not predict many flights are delayed. Most of delayed flights that it predicts are truly delayed, however, it also misses so many true delayed flights. That is why we have high precision and low recall. A figure below can explain this situation better.

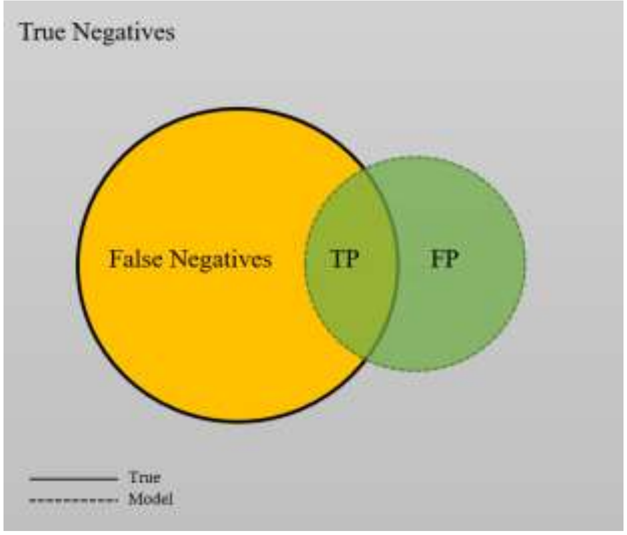

Figure 16: High Precision & Low Recall Scheme

Logistic Regression with Hyperparameter Tuning and Resampling

Before diving into what hyperparameters will be tuned, it is needed to recall how logistic regression works and what function it tries to maximize/minimize. The cost function of logistic regression that needs to be minimized is defined as:

$$\frac{1}{2}w^T w + C \sum_{i=1}^{n} \log\left(e^{-y_i(X_i^T w+c)+1}\right)$$

where C is a regularization parameter that penalize parameter space. This will be one parameter in our grid search that needs to be tuned. Other hyperparameter in grid search is class weights. This allows us to weight classes to specify their importance.

Since there is an imbalanced class distribution in our dataset, it is crucial to treat our model with respect to that. One way to solve this problem is applying a method in imbalanced-learn library. We chose to use random over sampler method due to computational expenses reasons. There are other great methods under that library, however, it comes with computational cost.

With these settings, means of the results are provided in a table below.

Performance Metrics	**Baseline**	**Logistic Regression**
Precision (Delayed Class)	0.80	0.70
Recall	0.50	0.74
F1-Score	0.61	0.72
AUC	0.83	0.87
AP	0.76	0.81
Training Accuracy	0.79	0.80
Testing Accuracy	0.79	0.80

Table 4: Results of Logistic Regression

Table shows that over sampling and hyperparameter tuning find a balance between precision and recall while keeping accuracy same. In this type of problem, it is good to have less false negatives than false positives, in other words, we want the model to focus more on recall than precision. Even though we tried to maximize F1 score, a harmonic mean of precision and recall, the average recall metric of the model is slightly higher than precision, which is what we wanted.

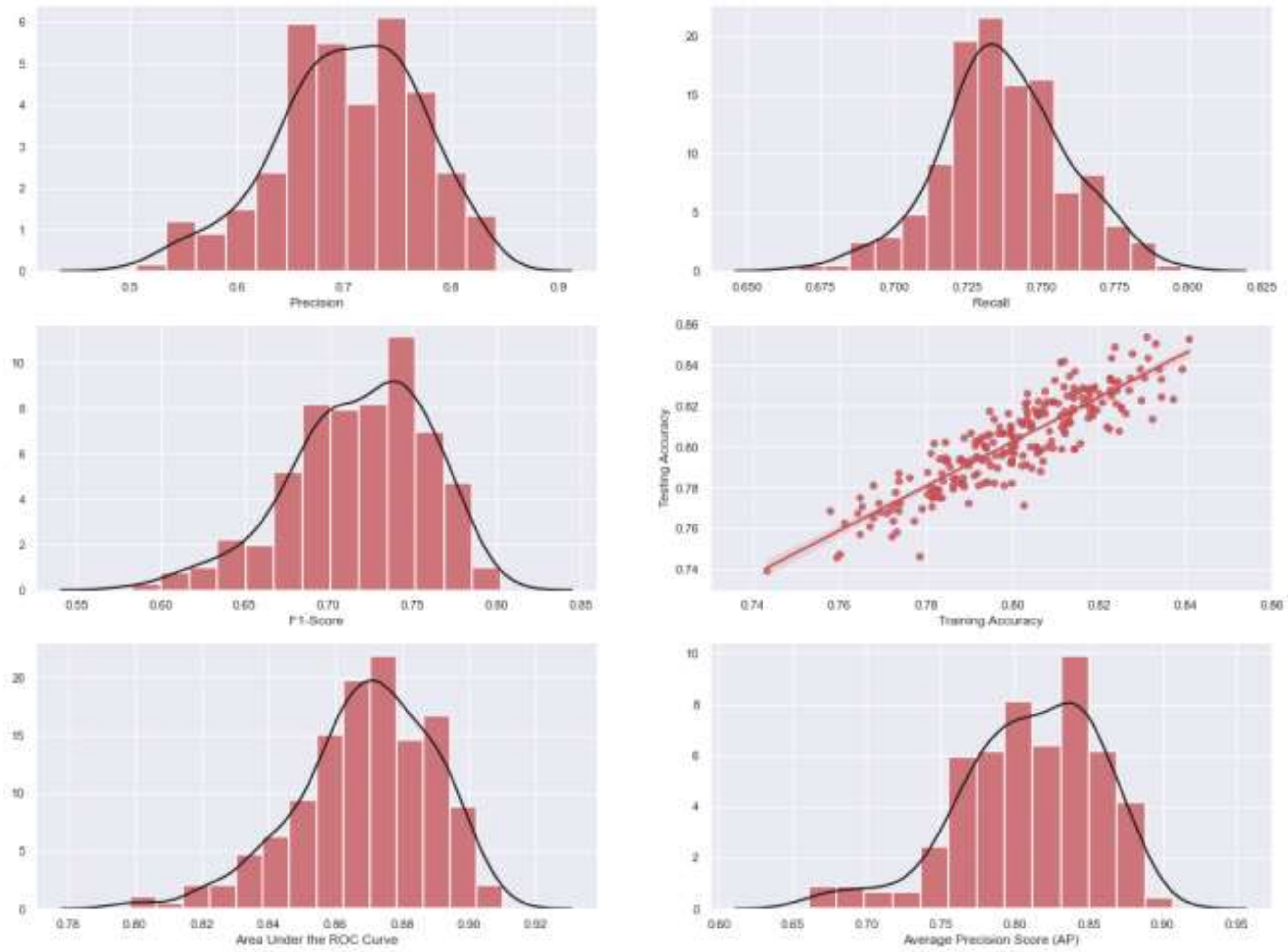

Figure 17: Histogram of Results from Logistic Regression

Conclusions and Future Work

Flight delays are a long-standing problem that has been studied previously mainly using historical records. In this work, we present an approach for flight delay prediction by using logistic regression model. Our earlier studies showed that taking late aircraft factor into an account improves our result significantly. That is why, we included this feature along with time difference between arrival time and next departure time of an aircraft. Since flight delays happen only 18-20% of the time in 2018, we need to treat it as an imbalanced problem. In this case, it would be helpful to import imbalanced-learn library. In our preliminary studies, we tried random under sampling, random over sampling, SMOTE, and SMOTEENN sampling techniques. All sampling techniques take significant amount of time but random over or under sampling techniques. We preferred to go with random over sampling due to loss of information. We were able to increase the performance of baseline model because we converted our given non-linear features into a larger number of linear features (Table 2: Model Features) and we over sampled data of minority class. In every algorithm we use, the performance on the training set significantly beats the performance on the test set. One way to diagnose the causes of this supposed over-fitting is to plot training set versus the testing set accuracy. We did not see anything significant data point below the regression line in Figure 17. The three major sources of improvement we implemented were: aggregation of several weather sources to get complete and accurate data, careful tuning of model's meta-parameters, feature engineering to capture time dependence. Our future work includes three different directions:

Generate new features that capture the nuances of the codependence between a flight's delay and the next flights delay (as of now, we only account for first-order time dependencies, but the interactions in reality are more complex).

Use additional features (twitter data, finer resolution terror data, etc.) to account for factors that contribute to delays but are invisible by our algorithms as of now due to the absence of proxy variables.

Use more of the already available data to improve our estimates on the delayed class.

Recommendations for the Client

There are two major sides that are affected from flight delays: clients and airlines. Delays cost money to airlines for an almost $8 billion a year aside from its cost to clients such as time, business losses, and money. If an airline company analyze underlying delay causes using machine learning algorithms, it can focus on reducing those causes, letting passengers know ahead of time, and optimization their flight operations and crew. From passenger's side, flight delay prediction models can provide probability of being delayed while they are booking it. In addition, it would be helpful to build an early alert system for passengers so that they can be notified before going to an airport.

References

[1] https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1
[2] https://www.transtats.bts.gov/Data_Elements.aspx?Data=1
[3] https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
[4] https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf
[5] Sud, V. P., Tanino, M., Wetherly, J., Brennan, M., Lehky, M., Howard, K., & Oiesen, R. (2009). Reducing flight delays through better traffic management. *Interfaces*, *39*(1), 35-45.