# Flight Cancellation Prediction Modeling

**Ying Li**

5/18/2023

# TABLE OF CONTENTS

# Introduction

The air cargo industry is a critical component of the US transportation and logistics infrastructure, providing a fast and efficient way to transport high-value, time-sensitive, and perishable goods. This industry comprises many players including but not limited to airlines, cargo carriers, freight forwarders, etc.  All the players work together to transport a wide variety of goods.  Over the last couple decades, the air cargo industry has been growing steadily, driven by the rise of e-commerce, globalization, advancement in technology, and improved logistics infrastructures.

As a leader in the U.S. air cargo industry, Quick Fly (QF) has been providing quality cargo delivery service for its customers throughout the country for the past decade.  Recently, it was brought to the attention of the leadership that QF has been paying higher than industry average claim payments for air cargo shipment delays caused by flight cancellation in the past 12 months, mostly for shipments departing Los Angeles International Airport (LAX).  QF needs to understand the root causes of the issue and turn the situation around with the ultimate goal to improve customer satisfaction, keep operation costs down and increase the overall profit.

The aim of this project is to use data science tools and techniques to evaluate and identify the elements that directly impact the likelihood of flight cancellation out of LAX, such as airlines, destination airports and departure days/time, etc.  The expected deliverable is a solid model that can accurately predict the likelihood of cancellation of outbound flights from LAX.

# Data

For the raw data, I choose the datasets of flight delays and cancellations published by the U.S. Department of Transportation (DOT) posted on Kaggle. The U.S. DOT Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

The datasets consist of 3 files.

- airlines.csv is the smallest file out of the three with 359 B. It contains 14 rows and 2 columns.
    - IATA_CODE → IATA airline codes
    - AIRLINE → Airline names
- airports.csv has a medium size of 23.87 KB. It contains 322 rows and 7 columns.
    - IATA_CODE → IATA airport codes
    - AIRPORT → Airport names
    - CITY → Airport cities
    - STATE → Airport states
    - COUNTRY → Airport countries
    - LATITUDE → Latitude of the airport
    - LONGITUDE → Longitude of the airport
- flights.csv is the largest file out of the three with a size of 592.41 MB. It has 5,819,079 entries and 31 columns.
    - YEAR → Year of the flight trip
    - MONTH → Month of the flight trip
    - DAY → Day of the flight trip
    - DAY_OF_WEEK → Day of week of the flight trip
    - AIRLINE → Airline identifier
    - FLIGHT_NUMBER → Flight identifier
    - TAIL_NUMBER → Aircraft identifier
    - ORIGIN_AIRPORT → Departing airport

- DESTINATION_AIRPORT → Destination airport
- SCHEDULED_DEPARTURE → Planned departure time
- DEPARTURE_TIME → WHEEL_OFF - TAXI_OUT
- DEPARTURE_DELAY → Total delay on departure
- TAXI_OUT → Time duration elapsed between departure from the origin airport gate and wheels off
- WHEELS_OFF → Time point that the aircraft's wheels leave the ground
- SCHEDULED_TIME → Planned time amount needed for the flight trip
- ELAPSED_TIME → AIR_TIME + TAXI_IN + TAXI_OUT
- AIR_TIME → Time duration between WHEELS_OFF and WHEELS_ON time
- DISTANCE → Distance between two airports
- WHEELS_ON → Time point that the aircraft's wheels touch on the ground
- TAXI_IN → Time duration elapsed between wheels-on and gate arrival at the destination airport
- SCHEDULED_ARRIVAL → Planned arrival time
- ARRIVAL_TIME → WHEELS_ON+TAXI_IN
- ARRIVAL_DELAY → ARRIVAL_TIME - SCHEDULED_ARRIVAL
- DIVERTED → Aircraft landed on airport that out of schedule
- CANCELLED → Flight Cancelled (1 = cancelled)
- CANCELLATION_REASON → Reason for Cancellation of flight:
  A - Airline/Carrier; B - Weather; C - National Air System; D - Security
- AIR_SYSTEM_DELAY → Delay caused by air system
- SECURITY_DELAY → Delay caused by security
- AIRLINE_DELAY → Delay caused by the airline
- LATE_AIRCRAFT_DELAY → Delay caused by aircraft
- WEATHER_DELAY → Delay caused by weather

# Data Wrangling

Capstone2_DataWrangling.ipynb

Out of the 3 datasets we have, airlines.csv and airports.csv are pretty straightforward and don't require much cleaning work. However, flights.csv is humongous and contains 5,819,079 entries with a lot of missing data. That's the file we spent most time cleaning and organizing. Here are the key issues identified with the flights.csv dataset and the approaches we took to address each of these issues.

## Missing data

1. All the delay reason columns including 'AIR_SYSTEM_DELAY', 'SECURITY_DELAY', 'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY' and 'WEATHER_DELAY', contain a significant number of missing values. If we are interested in investigating the reasons of flight delays, these columns will be of great use. However, the aim of our project is to find out what combination of elements could be used as an indicator of flight cancellation. Therefore, these delay reason columns are not relevant to our study and were dropped from the dataset.
2. The "DEPARTURE_TIME", "DEPARTURE_DELAY", "TAXI_OUT", "WHEELS_OFF", "ELAPSED_TIME", "AIR_TIME", "WHEELS_ON", "TAXI_IN", "ARRIVAL_TIME" and "ARRIVAL_DELAY" columns all have some missing data. Since the information contained in these columns relates to flights not cancelled, they were dropped.
3. The “DIVERTED" column has a lot of missing value. Since our study here is specifically about flight cancellation but not flight diversion, this column is dropped.
4. “CANCELLATION_REASON' column was also dropped after some consideration. This is another column with a large amount of missing data. While the information contained in the column (i.e., A - Airline/Carrier; B - Weather; C - National Air System; D – Security) is useful to investigate the causes of flight cancellations, it does not provide much value in the flight cancellation prediction model we were to build.
5. "TAIL_NUMBER" contains some missing data. Tail numbers are unique identification numbers for aircraft. Based upon common sense, tail numbers should not be a normal deciding factor in flight cancellation. Therefore, this column was dropped.

## Data Integrity

1. Since this project is focused on flights departing Los Angeles International Airport (LAX), we wanted to extract all entries with “ORIGIN_AIRPORT” column value as “LAX”. When attempting to do so, we found that in addition to the 3-letter airport code, there are a lot of 5-digit numbers and strings in the “ORIGIN_AIRPORT” column. First, we thought that these 5-digit numbers might be the airport zip codes. However, after research, that didn’t seem to be true, as the zip code for LAX is 90045, much bigger than any of the 5-digit numbers listed in the dataset. Further investigation led us to this Kaggle post, which addresses the cross-reference between each 5-digit code and its corresponding airport. With these additional airport code reference files, we were able to establish the cross-reference and drew the conclusion that both LAX and 12892 refer to Los Angeles International Airport., and successfully extracted all flights departing LAX.
2. The same issue exists for the “DESTINATION_AIRPORT” column. To address that, wefirst extracted all the destination airport codes that are 5-digit integers, then merged it with the airport code reference file to find out the corresponding 3-letter IATA code for each 5-digit airport code, and finally replaced all the 5-digit destination airport codes with their corresponding 3-letter IATA codes.

## Data Organization

1. The 'YEAR' column was dropped because all the entries have the same value, 2015.
2. The “ORIGIN_AIRPORT” column was also dropped because after above clean-up, all values in this column are the same, LAX.
3. After data cleaning, we merged and joined the modified flights file with airports.csv and airlines.csv respectively in order to include the destination airport city and state information from airport.csv and the airline name information from airline.csv in the final dataset.

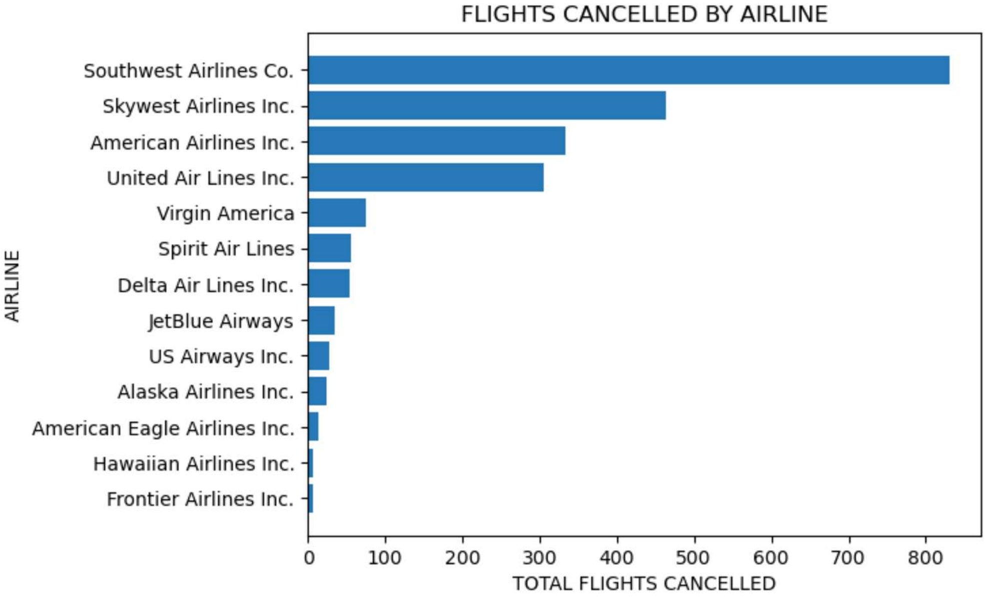
After data cleaning and wrangling, the clean organized data has 211388 rows and 13 rows including “MONTH”, “DAY”, “DAY_OF_WEEK”, “IATA_CODE, “FLIGHT_NUMBER”, “DESTINATION_AIRPORT”, “SCHEDULED_DEPARTURE”, “SCHEDULED_TIME”, “DISTANCE”, “SCHEDULED_ARRIVAL”, “CANCELLED”, “DESTINATION_CITY”, “DESTINATION_STATE” and “AIRLINE”.

# EDA (Exploratory Data Analysis)

Capstone2_EDA.ipynb

## Airline Analysis

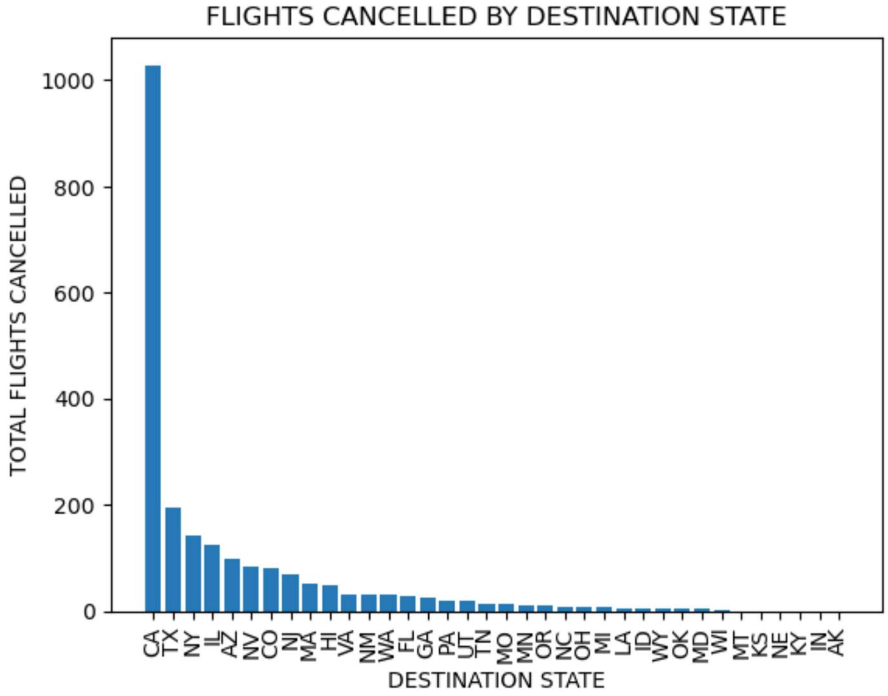
Initial analysis revealed that American Eagle Airlines ranked No.1 in terms of % of flights cancelled. However, after looking at the absolute number of flight cancellation, we found that the total of American Eagle Airlines flights cancelled was 13, extremely low. On the other hand, Southwest Airline had 830 flights cancelled, but its flight cancellation rate was very low at 2% because Southwest had a large number of outbound flights from LAX. Since we are interested in the absolute number of flights in this study, we decided to focus on the number of flights cancelled instead of cancellation percentages. Our analysis shows that Southwest, SkyWest, American Airlines & United Air Lines are the top 4 airlines that count for 87% of all the cancellations of flights out of LAX during the year.



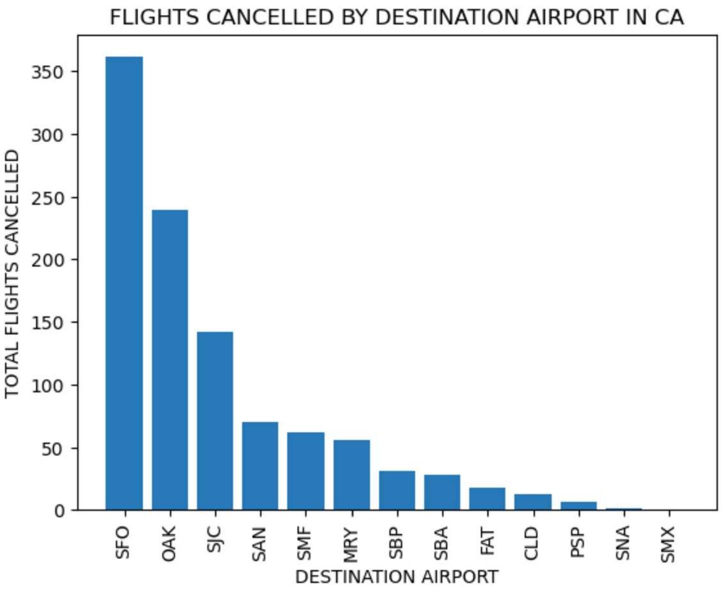
## Destination Analysis

Since the dataset contains a lot of destination airports, we decided to do an initial analysis based upon the states of destination airports. We found that CA ranks #1 for the total number of flights cancelled as well as percentage of cancellation. The 2nd place, Texas, has less than

1/5 of the total cancellation California has. Total flight cancellations for the other states are even lower, as illustrated in the bar chart.



Since CA has such a high cancellation number, we further looked into where in CA most of the flight cancellation occur.



It appears that most of the flight cancellation in CA occur with destination airports of SFO (San Francisco), OAK (Oakland) and SJC (San Jose). Interestingly, all three airports are in Northern

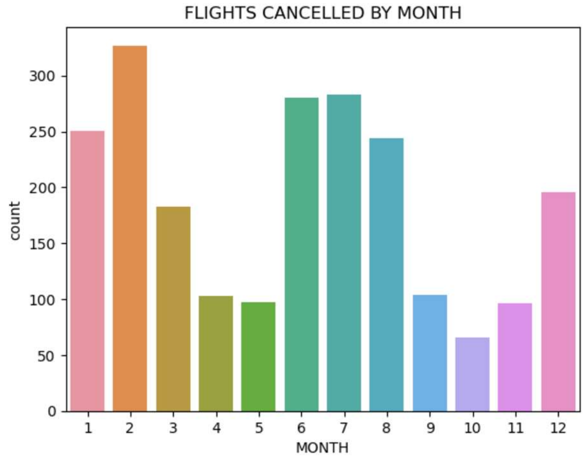
CA around the bay area. The flight cancellation numbers for these three destination airports are so high that they could have ranked number 2, 3 and 6 respectively out of all the states if these three airports were considered as individual states.

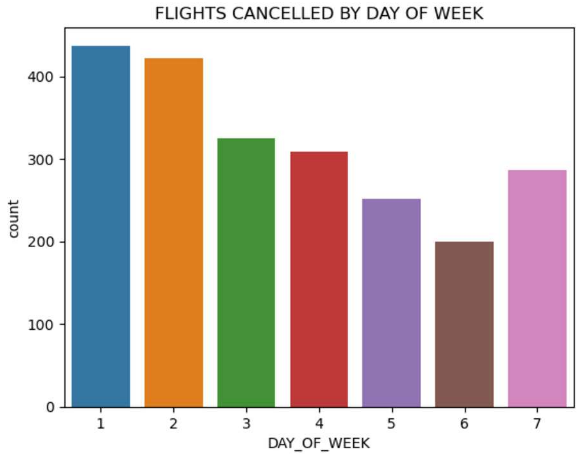
## Month, Day & Time Analysis

Next, we looked at the months, weeks and days and had some interesting findings.

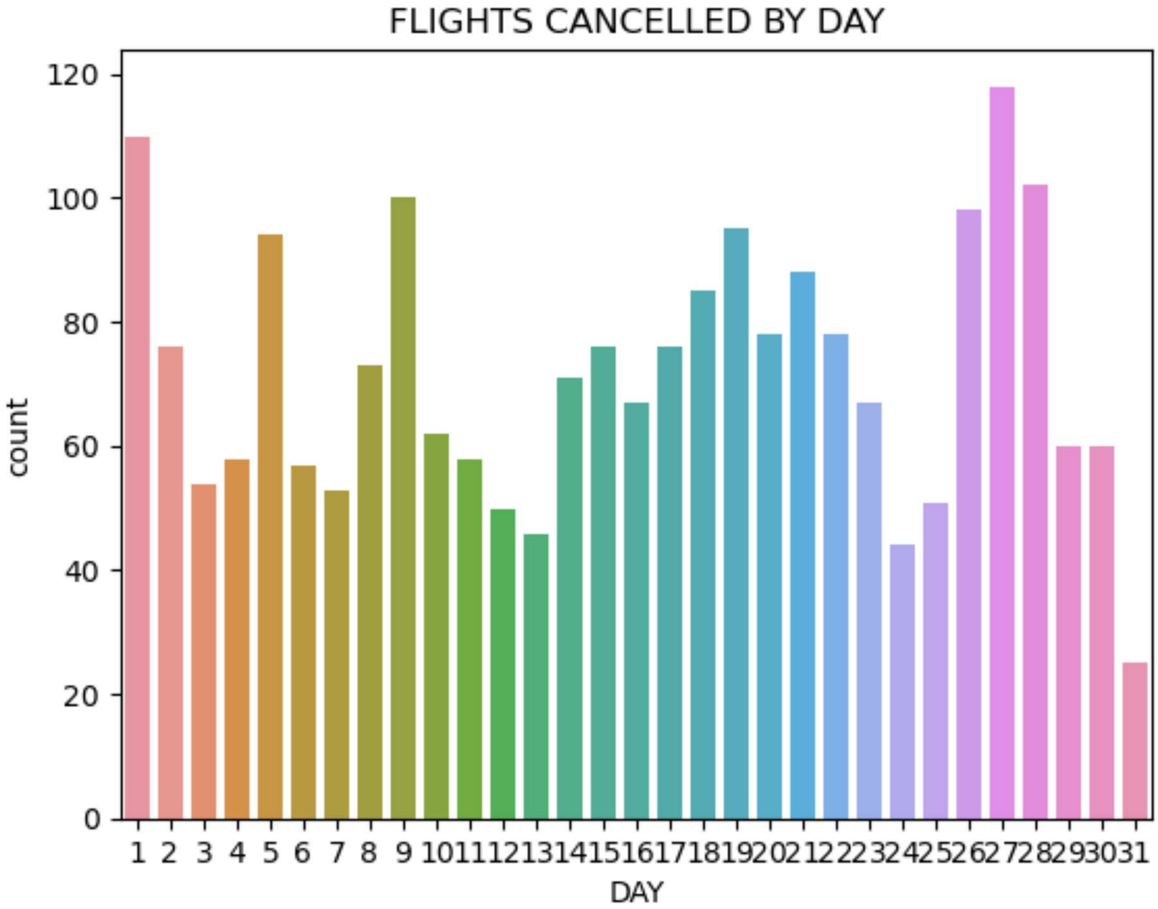
In terms of month, February seems to have the highest number of flight cancellations, 327. July and June are the 2nd and 3rd place, with 283 and 280 flight cancellations respectively. January and August numbers are also high, at 251 and 244 respectively. Overall, summertime and the beginning of the year seem to be the two periods with the highest flight cancellation numbers.

In terms of days of the week, Monday and Tuesday seem to have the highest number of cancellations. As the week goes by, the cancellation starts to fall gradually, and eventually rebounds on Sunday.





We also looked at the days of the month. Certain days during the month do have high cancellations, such as at the beginning of the month, toward the end but not at the end of the month, and around the middle of the month. However, there doesn't seem to be a clear pattern or obvious reasons showing why more cancellations occur on certain days of the month than others.

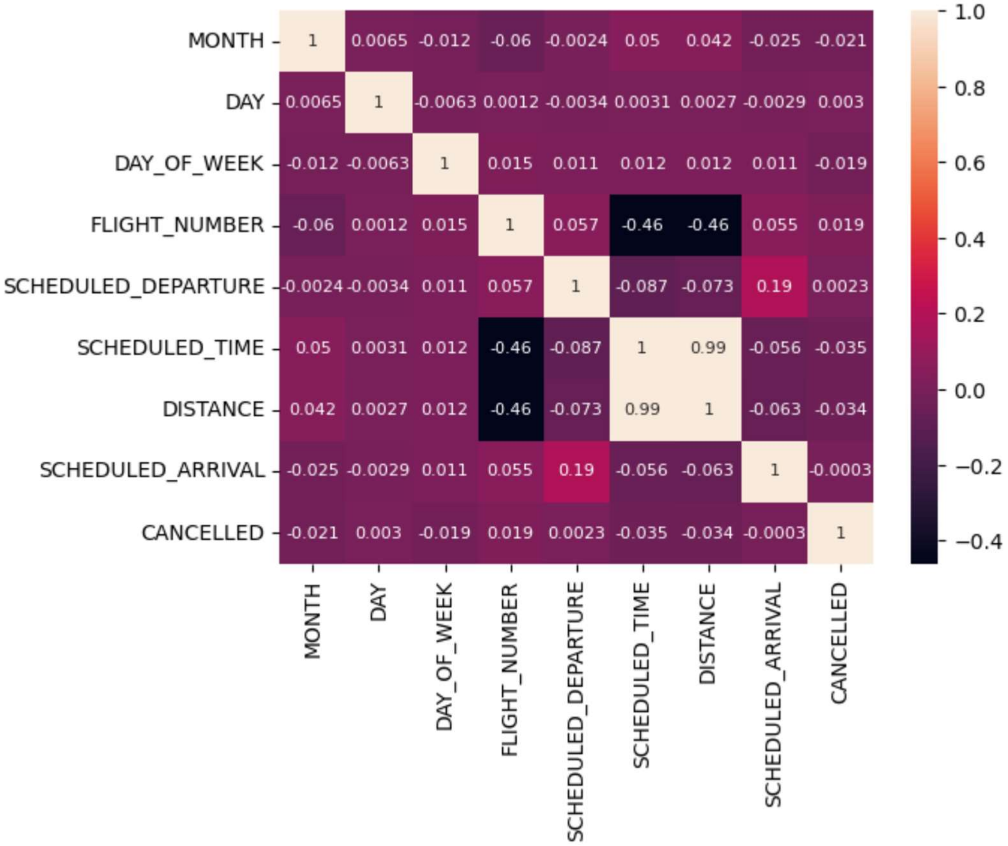


## Correlation Analysis

For the other features such as departure time, scheduled (air) time and arrival time, etc., since they all contain numeric values, we used a heatmap to evaluate the correlations these features have with flight cancellation.

As illustrated in the heatmap below, the scheduled time and the distance are highly correlated with the flight number. Also, no surprise, the scheduled arrival is correlated with scheduled departure.

The month, the day of the week, the flight number, the scheduled time and the distance all have a strong correlation with flight cancellation.  Scheduled departure and scheduled arrival appear to have the least correlation with flight cancellation, at 0.23% and 0.03% respectively. The day of the month also doesn't seem to be correlated with cancellation at 0.3%.



Based upon the EDA, we concluded that numerous features including “MONTH”, “DAY_OF_WEEK”, “FLIGHT_NUMBER”, “SCHEDULED_TIME”, “DISTANCE”, along with categorical features including “DESTINATION_AIRPORT”, “DESTINATION_STATE” and “AIRLINE”, all have relatively strong correlations with flight cancellation and will be our target features in the modeling phase of the project.

# Modeling

Capstone2 - Modeling.ipynb

Before modeling, we first label-coded all the categorical variables (i.e. "DESTINATION_AIRPORT", "DESTIANTION_STATE" and "AIRLINE") using LabelEncoder from sklearn.preprocessing.

We then standardized all the numeric features.  Histograms and boxplots revealed that none of the numeric features are normally distributed, therefore we used RobustScaler for “FLIGHT_NUMBER” which has a lot of outliners, and MinMaxScaler for all the other numeric features that don't have many outliners.

Another thing we did before we started modeling is to address the data imbalance issue.  Since only 1.05% of flights are cancelled during the year, we could tell that the dataset is extremely imbalanced. In order to optimize the modeling result, we used SMOTE to oversample the data in order to take care of the class imbalance issue before splitting the dataset into training and test sets.

Since the dependent variable, “CANCELLED”, is a categorical variable with 1 for cancellation and 0 for no cancellation, we decided to test out the following machine learning algorithms that are commonly used for classification problems.

1. Logistic Regression
2. k-Nearest Neighbors
3. Random Forests
4. Gradient Boosting
5. Extreme Gradient Boosting
6. Decision Trees

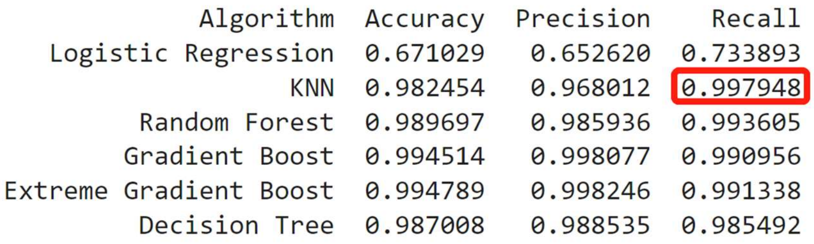
For each of these models, we followed the same process.

Step 1: We used grid search to find the optimal hyperparameters for each classifier. This is the most time-consuming part of the project.  Due to the large size of the dataset and the compound effect of trying various combinations of hyperparameters, sometimes it took us several hours to run the grid search and obtain the optimal hyperparameters.

Step 2: We fit the training data to the model and predicted the test data.

Step 3: We calculated the accuracy score, precision score and recall score, and printed out the confusion matrix and the precision report.

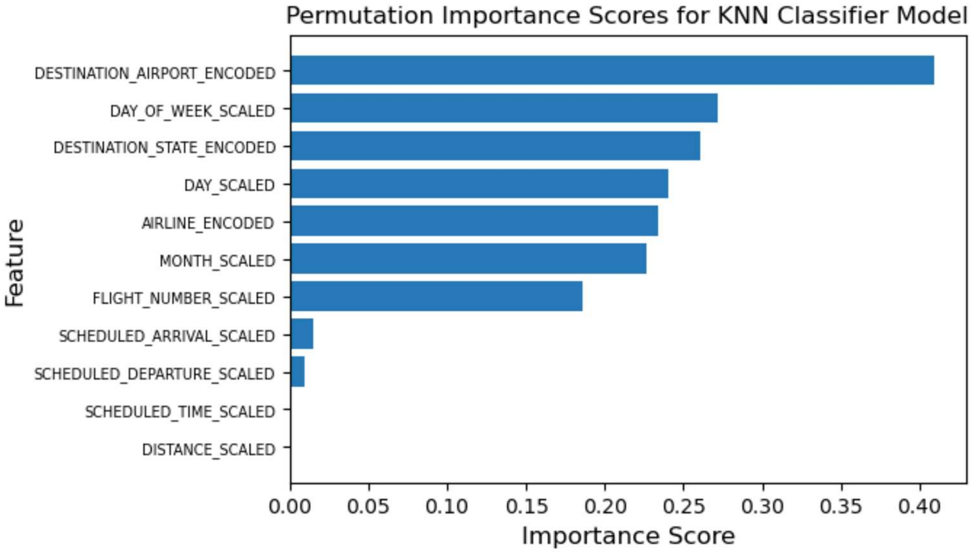
After we ran all 6 models, we compared the performance scores to pick the best algorithm. Since every flight cancelled means economic loss for the company, we want to identify and avoid as many potential flight cancellations as possible. Therefore, a high recall score is much more important than any other performance scores for our study.



```
              Algorithm  Accuracy  Precision    Recall
    Logistic Regression  0.671029   0.652620  0.733893
                    KNN  0.982454   0.968012  0.997948
          Random Forest  0.989697   0.985936  0.993605
         Gradient Boost  0.994514   0.998077  0.990956
 Extreme Gradient Boost  0.994789   0.998246  0.991338
          Decision Tree  0.987008   0.988535  0.985492
```

As illustrated in the above model metrics, the KNN model delivers the highest recall score of 99.7948%.  Also, the precision and accuracy scores from the KNN model are reasonably high in the 98% range.  Therefore, we conclude that KNN is our best performing model.

Since KNN is a distance-based algorithm, it doesn't learn a set of weights or coefficients that represent the importance of each feature. Instead, it calculates the distances between the query point and the training points and uses the labels of the k-nearest neighbors to make predictions. Therefore, the importance of each feature is implicitly determined by the distance metric used and the distribution of the data.  To get an idea of which features are most important for KNN, we used the feature permutation technique.



Two features, "DISTANCE_SCALED" and "SCHEDULED_TIME_SCALED", have extremely low importance scores.  That means they do not have much impact on the dependent variable.  To explore whether dropping these two features will get us a higher recall score, we took these two features out of the dataset, re-ran the grid search, and fit/tested the model. The performance scores we got from the updated model are 0.9974 for recall and 0.9814 for accuracy, very close to the original model.  Therefore, we decided to drop these two features.

# Conclusion

Based upon model metrics comparison, our best model is kNN, and our worst model is Logistics Regression.

For kNN, the following optimal hyperparameters are used:

- n_neighbors = 2
- p = 1
- weights = 'distance'

The kNN model has a recall score of 99.8%, which means it will accurately predict 99.8% of all flight cancellations.

Nine features are used in our model to predict flight cancellation.   The most important feature is destination airports, followed by days of week, destination states, days, airlines, months, and flight numbers.  Scheduled arrival and departure times also have an impact on flight cancellation, but not so strong as the other features.

# Future Improvements

We want to collect more data to improve the accuracy and reliability of this analysis.

1. The dataset used in this study is from Year 2015, which is quite outdated. Having newer data covering the last few years' flight information would give us a better and more up-to-date picture.
2. We want to obtain and identify new features from additional data, such as weather conditions, aircraft years in service, booking rates, etc., that could cast an impact on flight cancellation. This will enhance feature engineering and improve the overall outcome of the model.

We could also explore turning this into a regression modeling project by defining the dependent variable as the percentage of likelihood of flight cancellation and using a threshold such as 60% to identify flights that are likely to be cancelled.

By exploring the above points, we could improve the outcome of this project, achieve better accuracy, reliability, and effectiveness in drawing a conclusion on what combination of features are likely to cause a flight to be cancelled. With that understanding, Quick Fly will be able to identify and avoid booking those flights with a higher chance of being cancelled, and consequently lower the likelihood of paying late delivery claim charges.