Project Final Paper

Analysis of Airline Data

Mukund Kalantri
Computer Science
CU Boulder
Boulder, CO USA
muka4041@colorado.edu

Rohit Kharat
Mechanical Engineering
CU Boulder
Boulder, CO USA
rokh4336@colorado.edu

Reid Glaze
Mechanical Engineering
CU Boulder
Boulder, CO USA
regl1257@colorado.edu

ABSTRACT

In this project, we have airline data taken from approximately 2 million US domestic flights and analyze the delays and cancellations of these flights. We aimed to answer the following questions:

1) Which airports should be focused more on infrastructure development to mitigate the frequency and severity of delays?
2) What are the relationships between airline carriers and flight delays/cancellations?
3) What are the most common reasons for flight delays based on airports?
4) What are the relationships between days in a week or times in a day when there are a lot of cancellations and delays?

Additionally, we build a prediction model that will be used to calculate the arrival delay for a flight based on the other attributes. The customers and the airport authorities could directly use this. We also build several classification models to predict the reason for cancellations of flights based on the attributes provided by the dataset.

For our first question, we found out which airports we should focus on infrastructure development concerning carrier and aircraft delays. This information will help us to mitigate the frequency and severity of delays for passengers. The solution for the second question was found by which airlines have the most percentage of their flights delayed during departure and arrival based on different delay times. For the third question, we found the most common reasons for delays in flights based on origin airports, origin state, and airlines. The analysis was done to find the top 10 candidates which cause the delay in flights, so that information will help make informed decisions about which airlines and airports should be preferred. For our last question, we determined which times are optimal for consumers to purchase a flight to minimize the probability of delays.

We tried to build two types of prediction models. The first one was a Multiple Linear Regression model and the second one was a Neural Network. We concluded that the neural network was much better at predicting flight delays and the predicted values were much closer to the actual values. These predictions could be further improved with additional data and a more complex model.

Multi-classification models were built to classify or predict the reason for cancellations. The models which we built for this task were decision tree, random forest, AdaBoost classifier, extra trees classifier, gradient boosting classifier, XGBoost, Naive Bayes, and logistic regression. Among these models, we got the highest classification metrics values for the random forest classifier, followed by the extra trees classifier.

INTRODUCTION

For our first question, we find out which airports should be focused more on infrastructure development to mitigate the frequency and severity of delays. This would be helpful to airport authorities to improve

their services and it will also be helpful to the passengers to see which airport should be preferred while booking their flights.

The second question finds out what are the relationships between airline carriers and flight delays/cancellations. This will help the airlines to improve their services and passengers can make informed decisions about which airlines should be preferred.

The third question focus on the most common reasons for flight delays based on airports. This data would help airlines and airport authorities to improve their services which would be helpful to them in their business.

The fourth question tells us the relationships between days in a week or times in a day when there are a lot of cancellations and delays. This will help passengers to make an informed decision while booking flights concerning their timings.

Our classification model predicts the causes of delays, which helps the airport authorities in their data analysis. While the delay prediction model could be incorporated into websites and mobile applications to assist passengers in making their travel plans.

RELATED WORK

This dataset has been used for several different data mining tasks in a Coursera specialization course called "IBM Data Analyst":

1) Yearly number of flights canceled
2) Average delay time by airline
3) Monthly average delay for each type based on airline
4) Yearly number of flights delayed based on the arrival and departure states

The link to this specialization is:
https://www.coursera.org/professional-certificates/ibm-data-analyst

This dataset has also been used in another Coursera course called "Data Analysis with R" where it has been used for:

1) Performing Exploratory Data Analysis
2) Building a prediction model for predicting flight arrival delay
3) Using the R package tidymodels to evaluate the model

The link to this course is:
https://www.coursera.org/learn/data-analysis-with-r

DATA SET

We will be using a dataset called "Airline Reporting Carrier On-Time Performance Dataset" that contains information about US domestic flights that have occurred since 1987. This data was compiled from the Bureau of Transportation Statistics.

The link to the data set is:

https://developer.ibm.com/exchanges/data/all/airline/

Description of a few attributes is given below:

- Reporting_Airline: Airline Unique Carrier Code
- Origin: Origin Airport Code
- Dest: Destination Airport Code
- FlightDate: Date of Flight
- Cancelled: 1 = canceled
- CancellationCode: A = Carrier, B = Weather, C = National Air System, D = Security
- CarrierDelay: Carrier delay (minutes),

etc.

MAIN TECHNIQUES APPLIED

Data Integration

Since this dataset, we are using contains information for over 194 million flights and this amount of data is hard to work with, we are focusing exclusively on flights that have taken place in early 2022. Since it is

necessary to download each month separately, we have four different files containing flight data. We plan on combining these files to have one workable dataset. We will use the data for the months of January to April of 2022, which includes data for over 2 million US domestic flights.

Data Preprocessing

The initial data preprocessing steps carried out were importing required libraries and reading the collected data. We employed data cleaning techniques such as removing unnecessary features, handling missing data, and filtering outlier data. Delay attributes having null values were filled with zero because they indicated that the flights were canceled. Rows having missing values were removed because Data transformation methods were employed for transforming variables to correct data types for our computations. Numerical and categorical attributes were separated. The categorical attributes were encoded using one-hot encoding for avoiding the loss of contextual information caused by other methods, such as label encoding. However, label encoding was used for the target variable (i.e. cancellation code in case of classification models) which helped us to have a multi-class classification problem. The numerical values were checked for skew ed values with 0.75 as the limit and log transformation was applied to the attributes which exceeded this limit. This helped the attributes to get closer to Gaussian or Normal distribution. Also, attributes were checked for infinite values and appropriate elimination techniques were used to remove them. We standardized the data using sklearn's StandardScaler for reducing the effect of unscaled data. This data includes 109 attributes. Unnecessary attributes were removed to keep the most important features for our interesting questions. Correlation analysis was also used to remove features that exceeded 0.95 as the limit to avoid having redundant features. We engineered new features pertaining to different time periods for our analysis. We used the stratified shuffle data splitting technique for splitting the feature array and label array into training and testing sets. The test size was kept to 30% of the total data.

Data Analysis

We performed statistical analysis to describe our data and used pattern mining techniques to find answers to our interesting questions. Regression analysis will be done to estimate the relationship between the set of features. Correlation analysis was done to understand the effect of dependent variables on the independent variable. This helped us to perform classification as well as predictive analysis to identify future outcomes. Finally, time series analysis was used to identify trends, seasonality, and cyclic patterns in our data.

Data Visualization

We used data visualization techniques to display the findings of our project. We used trends, correlations maps, and bar graphs to show patterns in the data as well as for answering our interesting questions. For data reduction or showing important features, chart types such as bar graphs were useful. We used univariate analysis techniques such as distribution plots, box and whisker plots for outliers, and violin plots for kernel density estimation. In the bivariate analysis we used using line plots, bar plots, and scatter plots to show important patterns or clusters. We used Python data visualization libraries such as matplotlib, seaborn, and plotly for our visualization task.

Prediction

We build a delay prediction model where we tried two different approaches. It can be used to predict how long a given flight will be delayed. We did some data preprocessing to create our dataset where we filtered out the cancellation data and selected the attributes which are related to delays. We also made a train test split on our data in the 3:1 ratio. We used machine learning techniques like Multiple Linear Regression and Neural Networks and we evaluated our models

based on the mean absolute difference between the predicted and the actual values on our test set.

Classification

Several models were built for the multi-class classification task of classifying cancellation codes. This task helped us in identifying which attributes lead to the cancellation of the flight. Also, this helped us to understand the reason for cancellation based on the attribute values. The models which we built for this task were decision tree, random forest, AdaBoost classifier, extra trees classifier, gradient boosting classifier, XGBoost, Naive Bayes, and logistic regression.

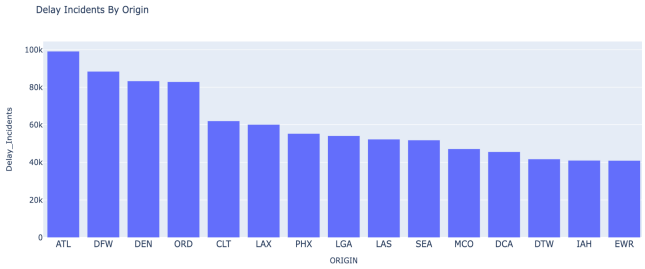
We got the highest classification metric on test data from the random forest classifier, which were 0.845 accuracies; 0.842 precision; 0.845 recall; and 0.841 F1 scores. ExtraTrees classifier performed second best among all the models with metrics around 0.83. Also, we used out-of-bag error for this classifier, to see how the model performed with an increasing number of trees. We found that the model performs wells with increasing trees with an out-of-bag error of around 0.175. The decision tree model gave good results as well as we were able to get information about the node counts and the depth of the tree. We used GridSearchCV for hyperparameter optimization on the AdaBoost classifier which gave fairly good results. For the GradientBoosting classifier, we fit the model with 15, 25, 50, 100, 200, and 400 trees and found that the error rate continually decreases. The metrics obtained from this classifier were around 0.80. XGBoost classifier performed averaged on this task with metrics values close to 0.76.

For the K-Nearest Neighbors model, we used the elbow method to determine the optimum number of neighbors and we got around 0.68 as the value for most of the metrics.We also built Naive Bayesian classification models such as Bernoulli, Gaussian, and Multinomial which gave metrics ranging from 0.15 to 0.66. The logistic regression model performed well and we got around 0.69 for the metrics values.

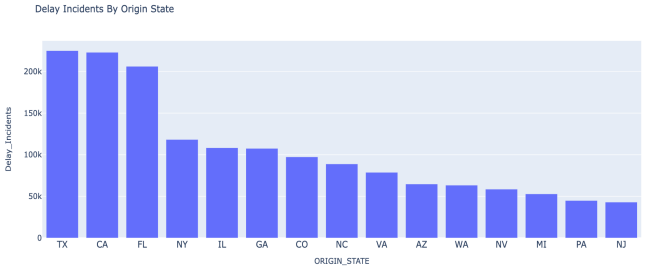
KEY RESULTS

Question 1

The information we wanted to seek from this interesting question was to understand which airports are affected by delays and how frequent the incidents happened at a particular airport and for a reporting airline. This information will be helpful in the future for infrastructure development considering delays are related to the airport. The analysis performed was helpful for us to understand which airports are frequently affected by delays and whether the type of delay can be used to mitigate that in the future.The figure below shows a bar graph for the top 15 airports having the most frequent delays:



The figure below shows a bar graph for the top 15 states having the most frequent delays:

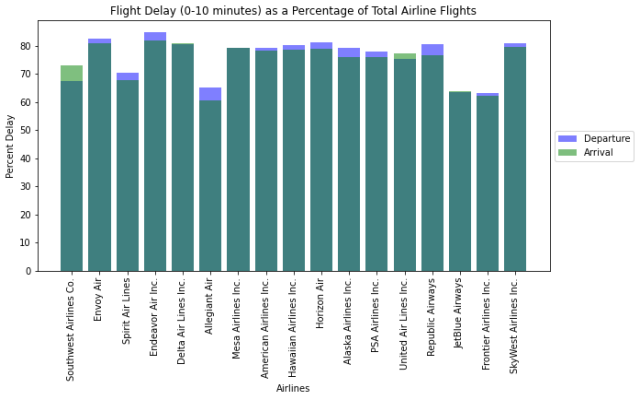


Question 2

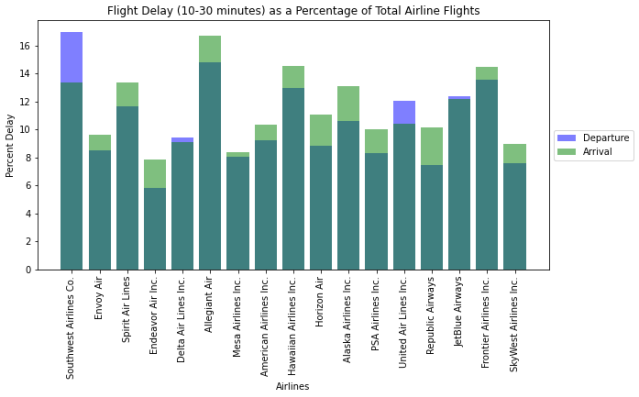
For this problem, we wanted to find those airlines which have a lot of delays and cancellations in their flights and which need to improve their services.

We filtered out our data based on arrival and departure for each airline first. The next step was to find out the unique airlines that we have. Since our original data only has airline codes, we found the airline names based on these codes from the Bureau of Statistics website. We then looped through all the airlines and

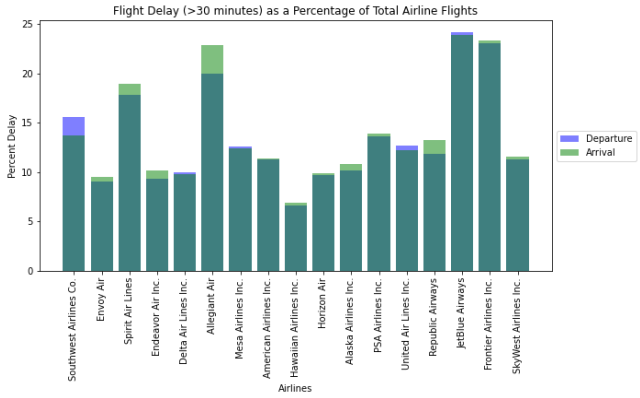
filtered out their data from the entire data we have. The next steps were to all those flights where there was no delay. Then for both arrival and departure cases, we removed those flights that had missing values in the arrival/departure columns. After this, we categorized all the flights with delays into three categories - Delay less than 10 minutes, Delays between 10 to 30 minutes, and Delays more than 30 minutes. Once we had these counts, we then took out the percentage that these delays were from the total number of flights for that airline. Our results are presented below.



We found that for all of the airlines, at least 60 percent of the flights had short (up to 10 minutes) delays. In these, Endeavor Air, SkyWest Airlines, and Republic Airways were the airlines with the most delays. Whereas, Delta Airlines, JetBlue Airlines, and Frontier Airlines had the best performance. However, the margin here is not much between these airlines, and based on the percentage of flight delays for all airlines, we can say that it is pretty normal for any flight to have short delays.

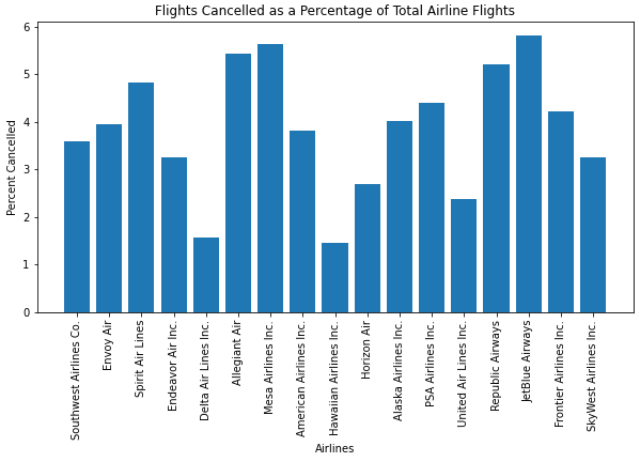


The above bar graph is for Medium length (10-30 minutes) flight delays. We found that in general, the arrival delays were more in comparison to departure delays, and since this happens with most airlines, we can say that the flight times mentioned while booking is not very accurate and is a little more than what is mentioned. We can also see that Southwest Airlines often get delayed even before leaving the airport so they should improve their pre-flight services (ticket checking, boarding luggage, aircraft systems check, etc). The best performing airlines in this category were Endeavor Airlines, Mesa Airlines, and SkyWest Airlines. Whereas the worst performers were, Allegiant Air, Frontier Airlines, and Hawaiian Airlines. We also saw that at least about 6 percent of all flights irrespective of the airlines get medium-length delays.



The long (more than 30 minutes) delays category is represented by the bar graph above. This is the category which should be focused on by the airlines to improve their services as such long delays give really bad experience to the customers and continued bad performance can cause problems to ticket sales for these airlines and affect their business. We observed that JetBlue Airlines, Frontier Airlines, and Allegiant Air have really bad performance here and more than 20 percent of their flights have long delays. The bar graph shows that there is a pretty big percentage gap between these airlines compared to their competitors, which should concern them and force them into doing detailed analysis to figure out the reasons for delays and improve their services. While the other airlines

had comparable performance here, Hawaiian airlines certainly have performed well which should be a big hope for them in terms of their ticket sales going up.

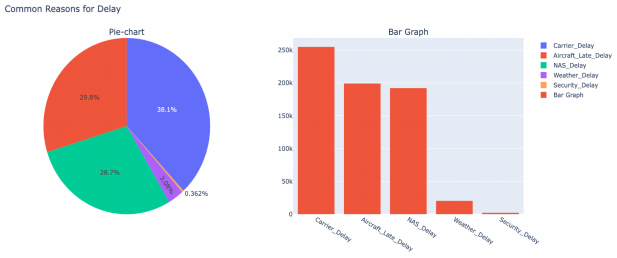


This last bar graph shows us the percentage of flights getting canceled by airlines. We can see that these percentages range from 2 to 6 percent of total airline flights in general. While JetBlue again turned out to be the worst performer, the other airlines with bad performance were Mesa Airlines and Allegiant Air. However, the difference between performance here for the remaining airlines does not vary much, it is worth mentioning that Hawaiian Airlines again had the best performance along with Delta Airlines, which is a big plus point for their business.

Question 3

The answer we were seeking from this question was what were the most common reasons for delays based on airport information. This information would be helpful for us to understand which airports are affected the most by a particular type of delay and to provide relevant information to airport authorities to work on improving their services. The analysis was further extended by considering the origin state of flights as well as which airlines have the most frequent cases of delays.

The figure below shows the most common reasons for delays:

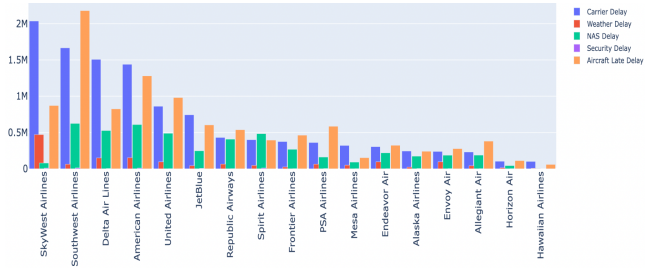


The figure below shows the origin airports most affected by carrier delays:

The figure below shows the origin airports most affected by security delays:

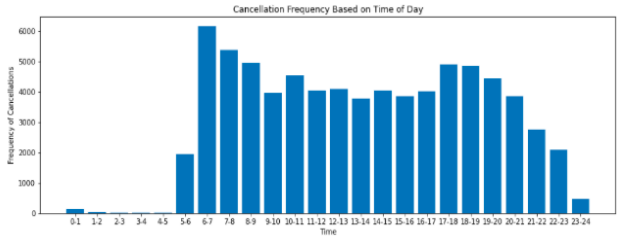
The figure below shows the origin airports most affected by late aircraft delays:

Further, the analysis was extended to understand which airlines were affected by a particular type of delay.

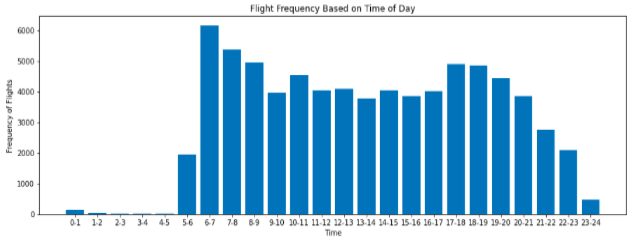


Question 4

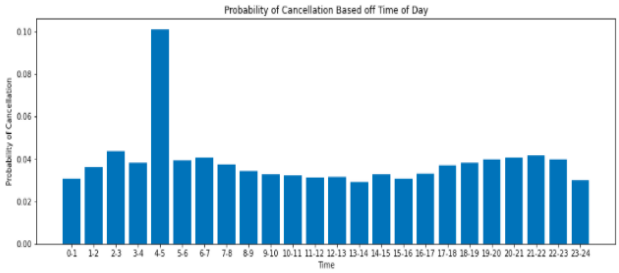
For this problem, we wanted to find information that would be useful to the consumer when booking a flight. Since our flight data only includes 4 months of the year, we decided to focus more specifically on times of the day. First, we focused exclusively on cancellations. We filtered out all the canceled flights into a data frame. Then we separated these flights into 24 different categories based on hourly intervals of departure times. This classification method resembles a decision tree.



The above bar graph shows that the highest frequency of cancellations occurs between 6 am and 7 am. However, this data is not particularly useful for the consumer because it does not give any information on probability. We performed the same type of classification for the data without filtering for cancellations and came up with a similar-looking graph.



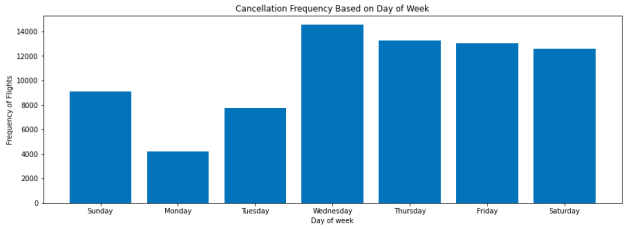
Next, we divided the frequency of cancellations for a canceled slot by the frequency of flights that departed in the same time slot. We came up with the following graph.



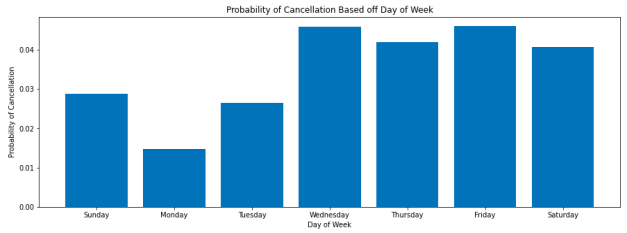
It appears that there is one significant outlier. Flights that departed between 4 am and 5 am had nearly a 10 percent chance of being canceled. It is unclear why this was the case. The rest of the distribution appears to be a lot more consistent, with the cancellation rate hovering between 2.9 and 4.4 percent. According to this data, a consumer should book a flight between 1-2 pm if they wish to avoid cancellations. The second lowest probability of cancellations occurs between 11 pm and 12 am, but this probability increases after 1 am and before 11 pm. Generally speaking, a consumer should aim to depart in the middle of the day and avoid the time slot between 4 am and 5 am if they wish to avoid cancellations.

In addition to the time of day, we set out to answer how the day of the week would affect the probability of cancellations. First, we found the frequency of flights that occurred on each day of the week. These are depicted in the graph below.

Next, we found the frequency of cancellations that occurred on each day of the week. These are depicted in the graph shown below.



Then we found the probability of cancellations occurring by dividing the frequency of cancellations by the frequency of flights. This is depicted below.

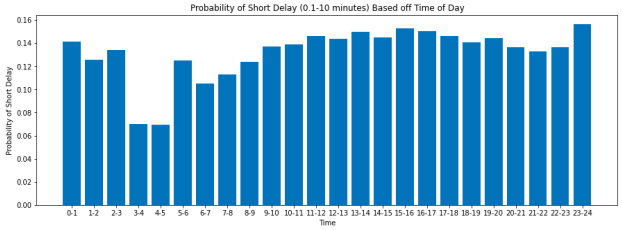


It appears that there seems to be a similar amount of flights operating on each day of the week. However, cancellations seem much more frequent later in the week. To find the probability of cancellation, we divided the cancellation data by the flight data.

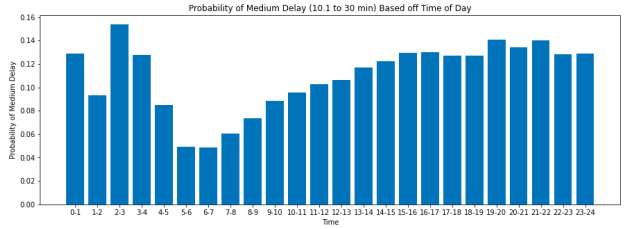
It appears that cancellations are most likely to happen from Wednesday to Saturday. They are least likely to happen on Monday. In summary, if someone wants to avoid getting their flight canceled, they should aim on flying on Monday, close to 1 pm.

Additionally, we analyzed the probability of a delay, separating delays into 3 different categories. We had short delays (0.1 to 10 minutes), medium delays (10.1 to 30 minutes), and long delays (30.1 to 60 minutes). For each category, we used the same strategy that we used for cancellations. First, we found the probability that the delay would occur in each time slot. Next, we divided this data by the total number of flights that occurred in each respective time slot. From this, we derived the probability of each delay type occurring in the given time slot.

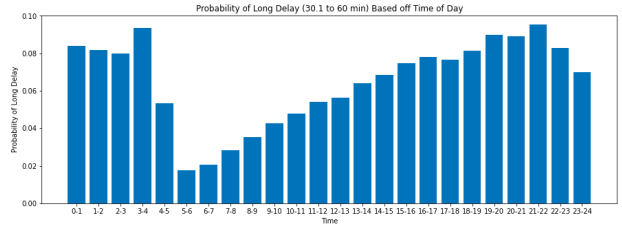
For short delays (0.1 to 10 minutes) we came up with the graph below.



For medium delays (10.1 to 30 minutes), we came up with the graph below.



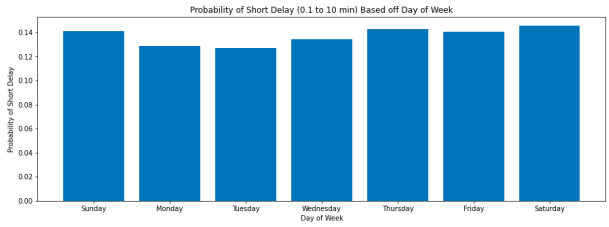
For long delays (30.1 to 60 minutes), we came up with the graph below.



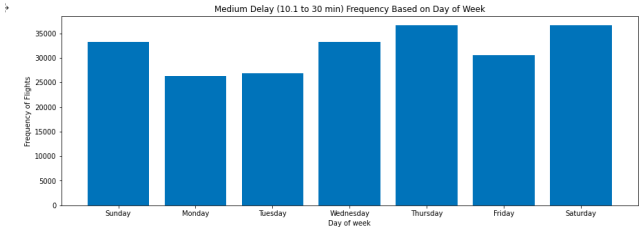
The patterns in the 3 graphs are rather consistent. However, the graph for longer delayed data frames exaggerated the difference in hour-by-hour distributions. In summary, if a consumer wishes to avoid delays they should aim to depart between 5 and 7 am and avoid departing late at night or very early in the morning before 4 am.

We performed a similar analysis for days of the week, analyzing the probability of short delays, medium delays, and long delays occurring each day.

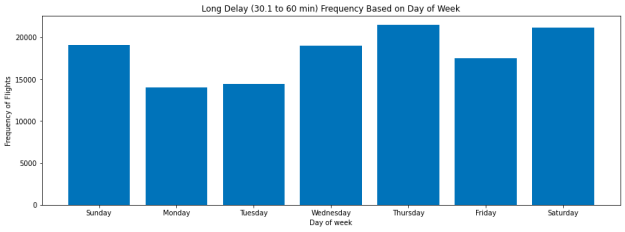
For short delays (0.1 to 10 minutes), we came up with the graph below.



For medium delays (10.1 to 30 minutes), we came up with the graph below.



For long delays (30.1 to 60 minutes), we came up with the graph below.

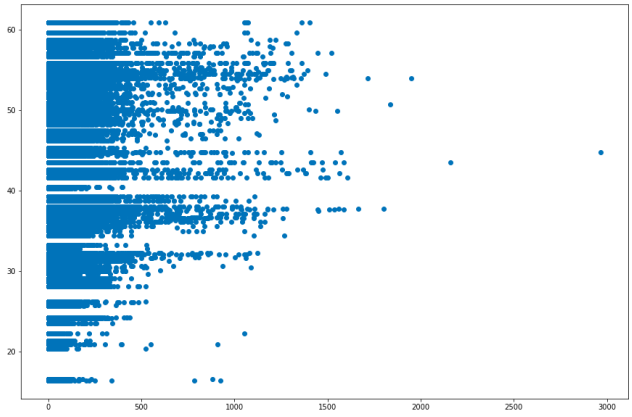


Similar to our graphs for delays based on departure time. the graphs for the three different delay types show consistent patterns but are more exaggerated in the long delay category. A consumer wishing to avoid delays should aim to depart Monday or Tuesday and avoid departing Thursday or Saturday.

From analyzing this delay data, we can also roughly determine the frequency of each type of delay. Short delays happen roughly 13 percent of the time. Medium delays happen roughly 10 percent of the time. Finally, long delays happen roughly 6 percent of the time. This does not include delays that are over 60 minutes long, but these are relatively uncommon. Of course, these are rough estimates and a more specific analysis would be recommended to get more accurate measures.

Prediction Model

For our regression model we tried two different approaches - for the first one, we set out to use MLR (multiple linear regression) to predict the delay time. We originally focused on 7 input attributes, 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'REPORTING_AIRLINE', ORIGIN_AIRPORT_ID', 'DEST_AIRPORT_ID', 'CRS_DEP_TIME', 'CRS_ARR_TIME'. We had 'DEP_DELAY' as the output. Since all of these input variables were categorical, we decided to use one hot encoding. This proved to be a challenge as we were forced to create a data frame with a ton of attributes and we were unable to fit a regression model. We reduced the attributes to two, 'DAY_OF_WEEK' and 'REPORTING_AIRLINE', and used one hot encoding.



The above scatter plot shows the actual values on the x-axis and the predicted values on the y-axis. Ideally, this plot should show data points on the y=x line for a good prediction model.

This prediction model was a failure and it can be concluded that day of the week and airline cannot be used exclusively to accurately predict the length of a delay.

The table below shows the difference between actual values and predicted values.

	Actual Value	Predicted value	Difference
0	5.0	42.152344	-37.152344
1	4.0	36.113281	-32.113281
2	65.0	38.750000	26.250000
3	11.0	31.753906	-20.753906
4	23.0	46.503906	-23.503906
5	1.0	29.105469	-28.105469
6	2.0	36.113281	-34.113281
7	9.0	37.980469	-28.980469
8	36.0	28.144531	7.855469
9	17.0	46.960938	-29.960938
10	17.0	53.660156	-36.660156

Then, for the second approach, we thought of using a neural network as we thought it would be able to figure out the relationship between all the different attributes that we initially set out with. So, we first filtered out all the records where flights were canceled, d of delayed, as they were of no use to our model. Then, we decided to use LabelEncoder() from Sklearn to encode the categorical values. instead of one-hot encoding. We did this to avoid the curse of dimensionality issue we had in the previous model. So, we converted the attributes 'DEST_AIRPORT_ID', 'ORIGIN_AIRPORT_ID', and 'REPORTING_AIRLINES'.We also decided to remove the minutes part from our attributes scheduled 'DEPT_TIME', and scheduled 'ARR_TIME'. This was done to reduce the overall number of categories in each of these attributes. After this, we split our dataset into train and test sets in the ratio of 75:25, and finally, we used StandardScaler() to normalize our dataset, and then we moved ahead to use the neural network. We used a simple Artificial Neural Network model with four Dense layers with nodes - 160, 480, 256, and 1 respectively. We also used a Dropout layer with a value of 0.2 after the first and second Dense layers. We also used MeanSquaredLogarithmicError() as our loss function and Adam optimizer with a learning rate of 0.01. This model was trained for 500 epochs. As we thought, this model turned out to be much better than the previous simple MLR model. The predictions were pretty close to the actual values. However, there were a few cases where the predicted values were very far from the actual values, which increased the overall mean absolute difference between the predicted and the actual results. This might be because the model was not complex enough to generalize well. However, with some more hyperparameter optimization, these few error cases should be easily handled in the future. Here are a few records from our results table:

	Predicted Delays	Actual Delays	Difference
0	2	0	2
1	2	2	0
2	0	0	0
3	0	0	0
4	1	0	1
5	0	0	0
6	-4277	0	4277
7	2	2	0
8	0	0	0
9	2	0	2

Although these results are very good, we believe that this type of delay prediction model could be made even better with the addition of some more attributes. In real life, flights are delayed due to many other reasons, for example - bad weather conditions, problems with the aircraft, security issues at the airport, air traffic, etc. In this dataset, we did not have the data related to these factors. We believe that with access to these attributes along with a more complex and better tuned neural, this type of delay prediction model could do well.

Classification Model

The classification model helped us in identifying which attributes lead to the cancellation of flights as well as we were able to build a multi-class model for the prediction of the reason for cancellation in the future. The model served the purpose of pattern mining for cancellations as we were able to understand the patterns leading to cancellation.

The models which we built for this task were decision tree, random forest, AdaBoost classifier, extra trees classifier, gradient boosting classifier, XGBoost, Naive Bayes, and logistic regression. We used several classification metrics for assessing the performance of the model. The metrics will be discussed while describing the model.

The first model we built was a decision tree classifier. The decision tree classifier performed well on the preprocessed data. We were able to get 0.83 accuracies, 0.76 precision, 0.73 recall, and 0.74 F1 scores. We also got the node counts and the depth of the tree from the model which were 15207 and 70 respectively.

	train	test
accuracy	0.979830	0.833624
precision	0.981415	0.764656
recall	0.964180	0.734179
f1	0.972465	0.748603

Next, we built a random classifier model which performed the best among all the models. Following is the image which shows the predicted labels and true labels for a sample array:

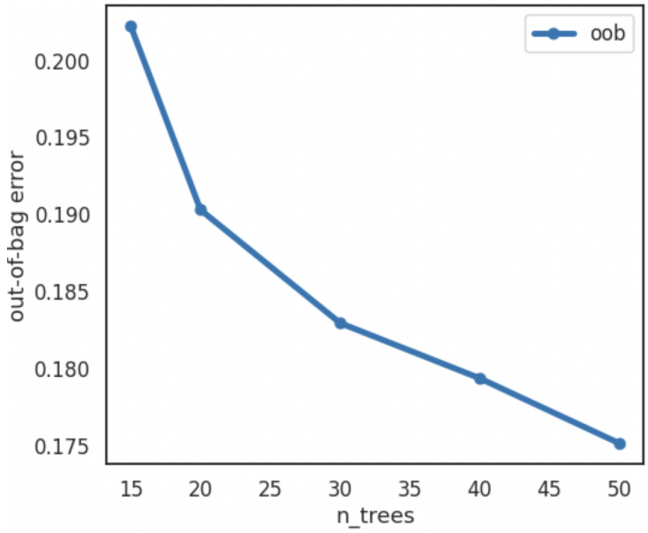
```
Prediction by Random Forest:  [1 1 1 0 1]

Actual Labels:
 2300     1
45004    1
16547    1
40609    0
36038    1
```

Below image shows the metrics:

	train	test
accuracy	0.979830	0.846911
precision	0.979836	0.843944
recall	0.979830	0.846911
f1	0.979808	0.843765

The second best performing model was the extra trees classifier and we used out-of-bag error to estimate the number of trees to be used in the model.



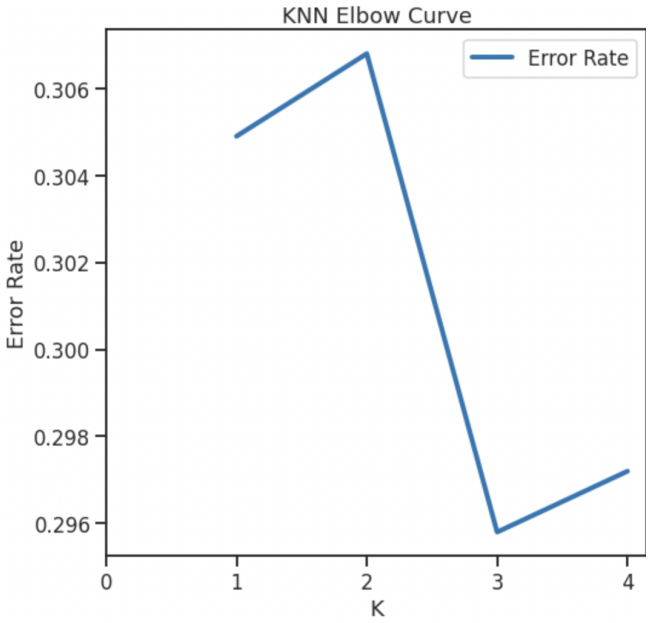
The following figure shows the metrics:

	train	test
accuracy	0.979734	0.834340
precision	0.979751	0.830827
recall	0.979734	0.834340
f1	0.979725	0.830264

Gradient boosting classifier also performed well on the dataset and gave around 0.80 metric value for accuracy, precision, recall, and F1 score. The model was also built on several trees and error was used to estimate the optimum number of trees.

XGBoost and AdaBoost classifier performed with 0.75 and 0.6 metrics values respectively.

For KNN, we used the elbow method to estimate the optimum number of neighbors.



However, with three neighbors, we got around 0.68 for classification metric values.

Next, we built the Naive Bayes model (Bernoulli, Gaussian, and Multinomial) for this task. However, we did not get significant results from this classifier.

The logistic regression model gave around 0.69 for classification metric values.

APPLICATIONS

Our project can help with a lot of real-world applications. Our findings related to the delay in flights because of airports will help the airports with poor performance improve their services. Having good performance would also help these airports in terms of business as the airlines and the passengers would prefer more to travel via these airports.

The results where we see the airline carrier performance in terms of delays will force the poor performing airlines to improve their services otherwise it would have a big impact on their business. These results also help the passengers to select better and more reliable flights.

We have also done an analysis of the reasons for delays at each airport which again would help the airport authorities to see what could be done to have better flight operations.

Our detailed analysis of delays based on the day of the week or time in a day would be more helpful to the customers when they are booking flights, as they can update their travel plans while considering the possibility of delays.

The classification model will serve the purpose of providing the reason for the cancellation of a flight to the airport authorities. This information can be shared by the authorities with the customer and can make an informed decision about the flight. This will also help in diverting a flight to a safer airport based on weather, national security, or any other issues. This knowledge will lead to help the airport as well as airline authorities to prevent losses incurred due to the cancellation of flights. Also, the customer will be provided with prior updates on the flight status.

The prediction model although not extremely efficient, gives decent predictions about the delays. Our experiments show how neural networks can be used in a more efficient way to make a very efficient and reliable delay predictor which could be eventually incorporated into airports' and airlines' websites and mobile applications. This would make it very helpful for the customers to chart out their travel plans with consideration of flight delays.