* + 1. **BIG DATA ANALYTICS**
    2. **(IST 718-M400: 37097)**
  1. **Homework Assignment:** Final Project
  2. **Student Name:** Sana Khan, Michael Morrey, Ryan Tervo
  3. .

**Introduction:**

One of the major areas of frustration in airline travel are late flights. For our project we attempted to predict flights taking off late using airline, airport, manpower, and airport weather data. The goal is that once a flight is predicted to be late then the airlines and airport can be notified and work to resolve the issue. Being notified to take action on flights likely to be late could help reduce delays and increase traveler’s overall flight experience.

**Summary of Findings:**

To give perspective of this modeling challenge, on average there are 45,000 flights across 5,000 airports each day. The Kaggle dataset from 2019 contain approximately 6.5 million flights with 28 variables, 15 variables were used to help predict late flights. In order to conduct our study 50K representative flights were selected randomly. This allowed the models to be trained and evaluated with the computing power available.

Using 15 predictor variables, six models were evaluated to determine the precision of the models. The best model discovered was Random Forest. While having perfect accuracy is always the goal, in this case the precision metric was considered to be more useful. This is because when a flight is to be identified as late then the airports and airlines can try to prevent the flight from being late. A false positive, in this case, takes resources from assisting flights that are truly in trouble of being late.

The trivial model showed that a randomly selected flight had an 18.9% chance of being late by 15 mins or more. A single model, Random Forest, was developed and was able to increase the precision score from 18.9% to 51.2%. This was a significant increase of 32.3%. However, it still meant that when a flight is predicted to be late it still only had just better than a 50/50 chance of being late. In order to improve the precision further, an alternative approach was developed.

The alternate approach from a single model was developed in which multiple customized models were created. The training data from these models was segmented by airline, airport, or a combination of airport and airline. Using this approach the precision score increased, in some cases to nearly 67%. While not every airline/airport combination could be evaluated (over 100k combinations), the combinations inspected consistently showed better results than 51.2%.

**Recommendations:**

Given the initial success of the multiple model technique, it is recommended that 25 Random Forest models be used over two years. That is, deploy the 25 models for two years and evaluate how they impact potential flight delays. If it can be shown that it decreases the number of late flights then this program should become a national wide early warning system. In addition, during those two years the models should be continually refined to increase their precision.

**SPECIFICATIONS:**

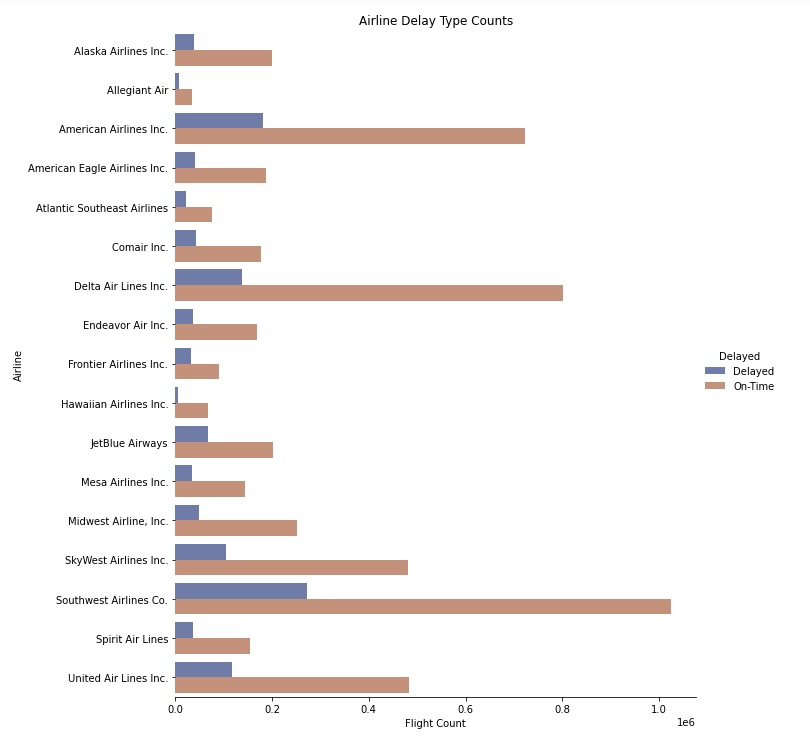
The U.S. aviation industry annual revenue is nearly $250 billion dollars. One major issue that plagues the aviation industry is delayed flights. The estimate is that nearly 20% of the flights, or one in five, depart late by at least 15 minutes. These delays are problematic for the airlines, airports, and passengers.

This data project works to establish an ‘early warning system’ which would give the airlines and airports an opportunity to intervene. If such a system could be established then the airlines and airports would have an opportunity to prevent a potentially delayed flights. Reducing late flights would save the airlines money, help with airport operations, and increase traveler satisfaction.

The dataset is from Kaggle and contains on-time and delayed flight data from 2019. It combines airline, airport, weather, and manpower data into a single .csv file. In total it is a 1.5GB dataset with nearly 6.5 million rows and 29 variables. There are 15 variables of interest that appear to be useful in predicting late flights. The 15 variables of interest can divide into four categories of manpower, airport/airline, time, and weather. The computational power was beyond what the team had access too. As such, a representative sample of 50K flights were selected to train using an 80/20 split.

**OBSERVATIONS:**

Figure 1- Airline On-Time and Delay Volume

Our initial observations of the data show that most flights are on time compared to delayed (Figure 1). Southwest Airlines, American Airlines, and Delta Airlines had the highest flight volume in 2019. These same airlines experienced the most delays, with Southwest having 27K delays, American Airlines with 18K delays and Delta with 14K delays (Figure 2).

Text

Description automatically generated

Figure 2 - Airline Delays

We then explored which airlines experience the most delays compared to total flights. The airline with the highest percentage of delays is Frontier Airlines, with over 25% of all flights resulting in a delay (Figure 3). JetBlue Airways and Atlantic Southeast Airlines followed with roughly 25% and 23%, respectively. Hawaiian Airlines had the lowest percentage by a considerable margin, with less than 10%. Delta Airlines, Midwest Airlines, and Alaska Airlines followed, with roughly 15% of flights resulting in delays.

Figure 3 - Airline Percentage of Flights Delayed

Chart, bar chart

Description automatically generated

We also investigated the delays per month. While there are no outliers, delays are higher in the summer months, the highest being in June (Figure 4). Another spike of delays occurred in December. This may correlate to when most people have time off to travel for summer vacation or the holidays. Conversely, the autumn months appear to have the lowest delays, with September being the lowest.

When evaluating percentage of flights delayed over time by airline, it appears that the airlines experience heavy delay times during different times of the year (Figure 5). For example, Hawaiian Airlines shows the lowest percentage of delays in the summer months and the highest in the winter months, whereas Frontier Airlines shows the highest percentage of delays in the summer months. This could correspond to the geographic location of where the airlines fly.

Figure 4 - Delays by Month

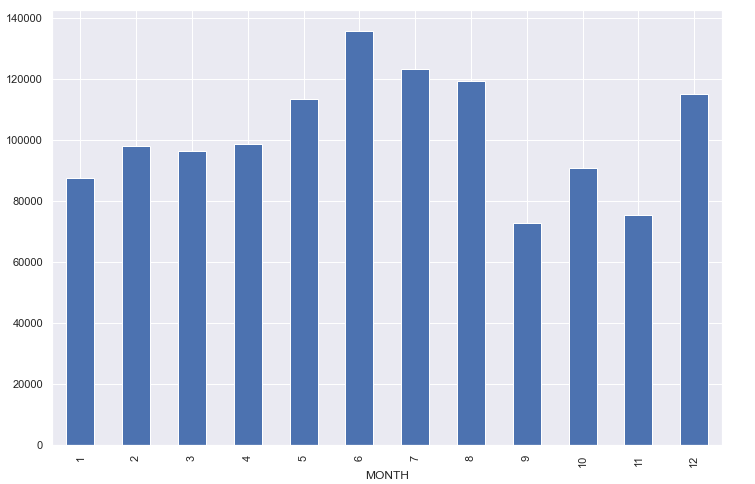


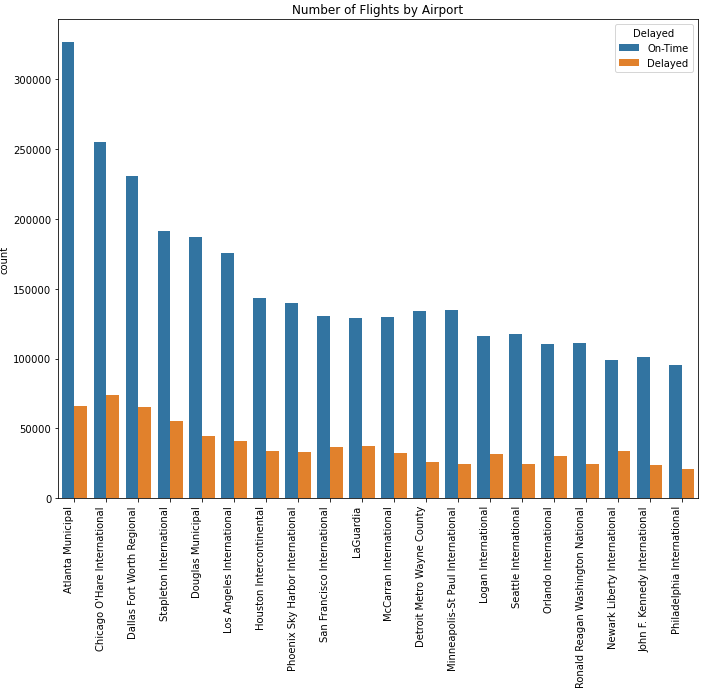
Figure 5 - Percentages of Flights Delayed by Month (Airline)

Chart, radar chart

Description automatically generated

When evaluating delays by airport, the airport with the highest volume of flights is Atlanta Municipal (Figure 6). Chicago O’Hare had the second highest volume but had more delayed flights than Atlanta. Dallas Forth Worth Regional had the third highest volume and delays.

Figure 6 - On-Time and Delay Volume by Airport



Next, we plotted the time blocks for the six most common airlines and observed that delays increase as the day goes on, with the highest delays occurring in the evening hours (Figure 7). The plot does show a few outliers, with Frontier having more delays than the other airlines in the red eye time frame, and Southwest having the highest delay from the 11:00 – 11:59 time block.

Figure 7 - Delay Percentages by Departure Time Block

Chart, bar chart

Description automatically generated

We also plotted delay percentage per day and observed that Wednesday had the highest percentage of delayed flights, with Saturday and Sunday having the second and third highest percentage (Figure 8). Monday and Friday had the least number of delays out of the entire week.

When evaluating the day of week delay percentages by the top 5 busiest airports, it appears that the heaviest delayed days can vary between airports (Figure 9). For example, the Chicago O’Hare International Airport experiences the heaviest delays on Saturday and Sunday. In contrast, Atlanta Municipal Airport has its heaviest delays on Wednesday and Thursday.

Figure 8 - Day of Week Delay Percentage

**Chart, bar chart

Description automatically generated**

Figure 9 - Day of Week Percentage by Airport

Chart, bar chart

Description automatically generated

Another component that was analyzed was variations in temperature, maximum wind speed, plane age, the number of flights that occurred during the same time at the airport (concurrent flights), and precipitation between on-time and delayed flights. Surprisingly, each of these variables had very similar variations between delayed and on-time flights. The box and whisker plots that indicate this are displayed in the appendix in figures 1a, 1b, 1c, 1d, 1e, 1f, and 1g. We expected each of these variables to have heavy influence on whether flights were delayed or not, but that does not appear to be the case. More relevant variables may have been discovered in the previous visuals, such as airport, airline, time of year, time of week, and time of day.

**Conclusions drawn from visuals:**

Upon reviewing the previous visuals, a passenger concerned about delays may want to avoid Frontier Airlines, JetBlue Airlines, and Atlantic Southeast Airlines, as these airlines tend to experience a very high volume of delays. Hawaiian Airlines, Delta Airlines, and Midwest Airlines experience far less delays and may be the better option. Passengers may also want to avoid traveling in the summer, which are the more heavily delayed months. Instead, spring or fall travel may be a better option. When considering airports, Atlanta Municipal and Chicago O’Hare International have a higher volume of flights, but they appear to be less likely to experience delays. Regarding time of day, red-eye and early morning flights may be less likely to see delays, although Frontier Airlines should be avoided during the red-eye hours. When considering the days of week to travel, the best days may vary depending on the airline. If traveling through the Atlanta Municipal, which is the busiest airport, a passenger may have the best chance of not experiencing a delayed flight on Sunday and Monday.

Additional visuals are also shown in the appendix that contributed to our analysis.

**ANALYSIS:**

Six models were developed to predict late flights. Those models included Logistical Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K Nearest Neighbor (KNN) and Naïve Bayes. Each model was evaluated on their precision score.

The precision measurement was chosen very carefully, because if and when a flight is identified as late, the airports and airlines can be notified and work to prevent the flight from being late. If a flight is misidentified as being late and is, in fact, a false positive then it costs the airlines and airports valuable resources to assist a flight that did not need it. As such, it is extremely important that flights identified as late are, in fact, likely to be late so that when efforts are made to prevent it from being late, it will have an actual positive impact.

The Logistic Regression, Support Vector Machine, and Naïve Bayes models predicted that all flights would be on time. The remaining four models predicted at least some late flights and produced precision scores. Their precision score ranged from 27.2% to 51.2%. These results can be seen in Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Logistic Regression | Support Vector Machine | Naïve Bayes | Decision Tree | K Nearest Neighbors | Random Forest |
| **Precision** | 0.0% | 0.0% | 0.0% | 27.2% | 32.7% | 51.2% |

Table 1: Initial Single Model (All Data) Precision Scores

In working to increase the model precision, an alternate strategy was developed and tested. The thought was that perhaps the level of complexity was beyond the model’s ability to accurately predict. As such, the solution of multiple models was developed to determine if the precision could be increased using different subsets of the data. In this case, the data was segmented by airline and airport. The models were trained and predicted on their respective data segments to determine if their performance increased beyond that of a single model.

This approach was applied on six data segments in which the models were trained and evaluated. In 80% of the test cases this technique outperformed the single model solution of 51.2%. These results can be seen in table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Set | Naïve Bayes | Decision Tree | K Nearest Neighbors | Random Forest |
| **1** | 27.7% | 29.6% | 45.3% | 67.5% |
| **2** | 27.7% | 29.8% | 45.3% | 65.3% |
| **3** | 20.0% | 22.1% | 29.4% | 49.5% |
| **4** | 42% | 38.0% | 53.2% | 64.6% |
| **5** | 0.0% | 25.5% | 36.8% | 62.7% |

Table 2: Multiple Models (Segmented Data) Precision Scores

Using multiple models appears to be a very good approach. In some cases, the precision score increased from 51.2% to 67.5%. In all but one case the model performed better. Only in the third run did the model’s precision decrease below 51.2% down to 49.5%.

**RECOMMENDATION:** What should your customer do?

The multiple model technique using the Random Forest model with segmented data worked very well. It outperformed the single model in 80% of the trials. It is recommended that this technique be applied to the top 20 airports and top 20 airlines for a period of two years. During those two years the models (400 of them) should be continually refined to increase their precision scores and data collected to determine how effective the ‘warning system’ was at reducing late flights.

At the end of the two-year trial the overall effectiveness of this program should be evaluated. If late flights can be predicted early and reduced, then the program should be implemented nationally. If the program was not effective at the end of two years, an assessment should be made to either to refine the process or conclude it.

**REFERENCES:**

The link to the dataset can be found here: https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations

**APPENDICES:**

Figure 1a - Max Temperature Variations

Chart, bar chart, box and whisker chart

Description automatically generated

Figure 1b - Max Wind Speed Variations

Chart, bar chart, box and whisker chart

Description automatically generated

Figure 1c - Plane Age Variations

Chart, box and whisker chart

Description automatically generated

Figure 1d - Concurrent Flight Variations

Chart, box and whisker chart

Description automatically generated

Figure 1e - Snow on Ground Variations

Graphical user interface, table

Description automatically generated

Figure 1f - Total Precipiation Inches Variations

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1g - Snow on Ground Variations

Graphical user interface, application

Description automatically generated

Figure 2a - Percentage of Total and Delayed Flights over 2019

Chart, line chart

Description automatically generated

Figure 2b - Percentage of Flights Delayed by Month

Chart, bar chart

Description automatically generated

Figure 2c - Airline Percentage of Flights Delayed by Month

Chart, bar chart

Description automatically generated

Figure 3a - Histogram of Delayed and On-Time Flights

Chart, histogram

Description automatically generated