Flight Delay Prediction and Analysis

Avinash Maharaj
University of Colorado Boulder
Computer Science
Graduate Student(In class CSCI 5502)
avinash.ratnavel@colorado.edu

SID: 107143926

Chirag Kamat
University of Colorado Boulder
Computer Science
Graduate Student(In class CSCI 5502)
chirag.kamat@colorado.edu

SID: 107164851

Rakesh Shivanand Margoor
University of Colorado Boulder
Computer Science
Graduate Student(In class CSCI 5502)
rakesh.margoor@colorado.edu

SID: 106964637

ABSTRACT

Use of flights for transportation is increasing day by day. Flight delays cause various problems to the passengers. It’s better if there is a system that predicts the likelihood of a flight being delayed by a certain time unit based on the behavior of the same flight in the past. That way passengers choosing to travel on that flight can be aware of its arrival time and hence those passengers can decide their itinerary accordingly. In this project, the idea is to build a similar system where we can generate the likelihood of flight delays from the past flight delay data and to find the reasons that causes such a delay. The reasons for the delay can vary based on the destination airport, region where the airport is situated, weather conditions, day of the week, month of the year, flight maintenance to name a few. Several data mining techniques such as logistic regression, Classification, Supervised learning are applied on past data to find interesting patterns in flight delays. This analysis will allow us to find the main factors that affect flight delays and predict the delay. The prediction model can be made more robust on past weather information as well.

CCS CONCEPTS

• **Data Mining**→ Correlation and Pattern Analysis;

• **Machine Learning** → Regression and Classification

KEYWORDS

Logistic regression, Bayesian networks, Classification, Correlation, Pattern matching, Supervised Learning, Random Forests

1 INTRODUCTION

1.1 Motivation

With the great increase in air traffic comes a large increase in the demand for airport capacity. These days, we are seeing a spike in the number of people travelling by flight. Airport capacity cannot keep increasing at a rate necessary to match the rising demand. When an airport's capacity is reduced during “peak hours", the demand for an airport's resources exceeds the capacity that the airport can afford. This is known as a capacity-demand imbalance. Demand refers to the number of flights scheduled to arrive or depart in a given time period (rate of flight arrivals or departures). Capacity is the maximum number of flight arrivals or departures in a given time period. The direct result of the capacity-demand imbalance is the airport congestion and flight delay.

Flight delays in the United States result in significant costs to airlines, passengers and society. These flight delays result in caustic disturbances to the aviation industry and its stakeholders. Given the uncertainty of their occurrence, passengers usually plan to travel many hours before their appointments, increasing their trip costs, to ensure their arrival on time. Furthermore, airlines suffer penalties, fines and additional operation costs, such as crew and aircrafts retentions in airports due to these delays. In 2010, FAA/Nextor completed a comprehensive study on the costs and impacts of flight delays in the U.S. and estimated the annual costs of delays in 2007 to be $31 billion. On the other hand, delays may also cause environmental damage by increasing fuel consumption and gas emissions. Such high delay costs and disturbance motivate the analysis and prediction of air traffic delays, and the development of better delay management model.

1.2 Background and Literature Survey

Flight delays can occur due to a variety of factors which can be more of less categorized into three categories. Factors associated with the destination airport, origin airport or in mid-air. Delays can also be related to airlines, and can also reactionary. Reactionary delays occur due to late arrival of previous flights.

Two government agencies keep air traffic delay statistics in the United States. The Bureau of Transportation Statistics (BTS) compiles delay data for the benefit of passengers. They define a delayed flight when the aircraft fails to release its parking brake less than 15 minutes after the scheduled departure time. The FAA is more interested in delays indicating surface movement inefficiencies and will record a delay when an aircraft requires 15 minutes or longer over the standard taxi-out or taxi-in time (Mueller, et al., 2002). Generally, flight delays are the responsibility of the airline. Each airline has a certain number of hourly arrivals and departures allotted per airport. If the airline is not able to get all its scheduled flights in or out each hour, then representatives of the airline will 8 determine which flights to delay and which flights to cancel (from http://www.travelforecast.com).

These delays take one of three forms, ground delay programs, ground stops, and general airport delays. When the arrival demand of an airport is greater than the determined capacity of the airport, then a ground delay program may be instituted. The airport capacity is unique to each airport, given the same weather conditions. The various facilities at an airport can determine how much traffic an airport can handle during any given weather event. Generally, ground delay programs are issued when inclement weather is expected to last for a significant period. These programs limit the number of aircraft that can land at an affected airport. Because demand is greater than the aircraft arrival capacity, flight delays will result.

There can be ground delay due to the popularity of the flight which can be greater than There are many studies that attempt to model or predict flight delays using data on origin, destination, time, and other non-weather features. These studies generally try to showcase some new theoretical work considering flight delay prediction as a model problem. In the last decade, there have been several attempts made to understand the weather's impact on the flight delays. Various parameters such as temperature, wind and precipitation are being considered to understand the weather's impact in a more accurate way.

Juan Jose Rebollo and Hamsa Balakrishnan [1] proposed a new class of models that considers both temporal and spatial delay states as explanatory variables. Random Forest Algorithm was used to predict departure delays. In addition to local delay variables that describe the arrival or departure delay states of the most influential airports and links (origin-destination pairs) in the network, the paper also proposes new network delay variables that characterize the global delay state of the entire National Airspace System at the time of prediction.

Young Jin Kim and Sun Choi [2] proposed a deep learning based approach for flight delay predication. The paper proposes an accurate and robust prediction model by combining multiple models based on the deep learning paradigm. Based on their evaluation, the Recurrent Neural Networks (RNN) has shown great accuracy in modeling sequential data.

2 PROPOSED WORK

2.1 DATA

The Bureau of Transportation Statistics has collected the US flight delay details for each year. Data has been downloaded from year 2012-2015. For each year there are around 4 million rows where each row corresponds to flight delay details such as origin airport, destination airport, arrival and departure time, weather delay, departure delay, arrival delay, airline delay.

2.2 Scope

We divide our scope into 3 major subdivisions. At each step we create a model by training it with the data from the year 2012-2014. This model would be tested against the available data of 2015. With this, we compare our model performance and accuracy which helps in figuring out any interesting patterns or correlation among the data.

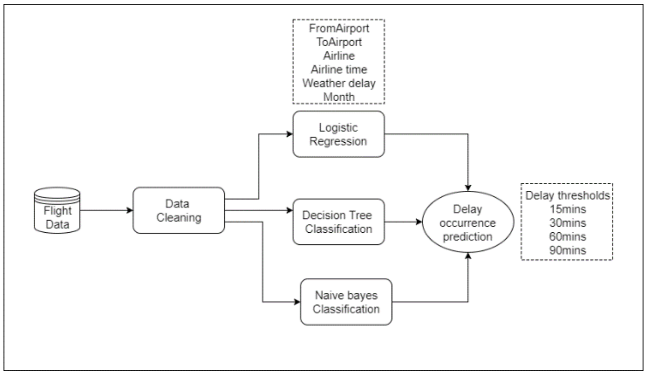


Figure 1: A basic flow diagram to illustrate the flight delay prediction model

2.2.1 Prediction based on Flight Data Only

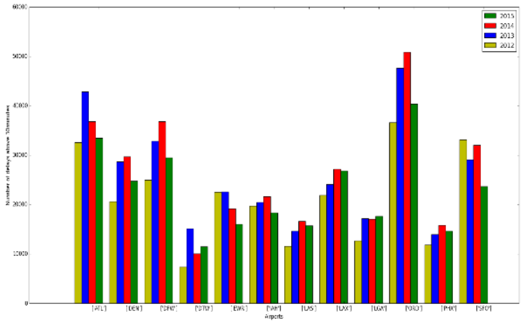


Figure 2: Flights delayed beyond 30 minutes for years from 2012 to 2015 for various airports

The basic task would be to create a simple flight delay prediction with the data that a user generally inputs while booking a flight. This includes the data related to origin-Airport, destination-airport, time of departure, time of arrival and airline. This would not consider the cost factor while predicting the delay. A logistic regression model is implemented by training on the above data and its accuracy is calculated through test data. The impact of individual parameters is analyzed through Bayesian networks. Figure 1 depicts basic analysis performed on different classifiers.

2.2.2 Prediction based on Flight and Weather data

With the basic delay prediction model built previously, we try to improvise it by making it more robust. The robustness can be achieved by training the model with the weather data that would impact the flight delay. Temperature, precipitation, wind and snow data is used to train along with the previous flight input data. The performance of this model is compared with the previous to understand the impact of weather on the flight delay problem. The Figure 1 depicts the flight delay models and the process involved.

2.2.3 Delay propagation analysis

Based on the prediction model built, now when a user would want to book the tickets, we would be able to predict whether there would be a delay and the time of delay on his booking. If, there is delay, the next approach would be to suggest alternative routes for the same Source and Destination with minimal deviation in the specified timings. With this approach, cost plays a vital role which is still ambiguous currently, as we do not have sufficient data to consider cost while suggesting routes. So currently, the alternate routes do not consider the cost factor.

3 EVALUATION AND RESULTS

3.1 Dataset evaluation

We started our analysis by considering various airports and how they are placed in terms of delays. For practicality purposes, a flight is considered delayed if its delayed by 30 minutes. The years in consideration are 2012 to 2015. We considered top 12 highest delayed airports as shown in Figure 2. We found out that Chicago's O'Hare International Airport had the most number of delays for all the 4 years. We next shifted our focus to airlines and delays associated with them. Figure 3 shows various delays occurring due to different airlines for the year 2015.

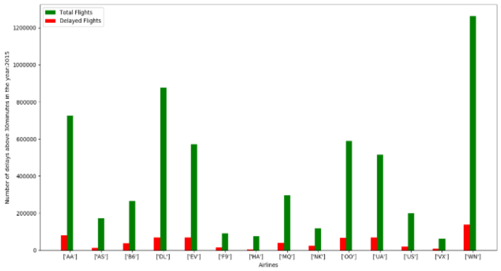


Figure 3: Flights that are delayed beyond 30 minutes for the year 2015

From Figure 3, we see that EV air-corporation flight has the highest percentage of flight delays in the year 2015.

We then focused on each month and collected the number of delays for years 2012, 2013, 2014 and 2015. From Figure 4, we found out that June, July and December were the most affected months in that order.

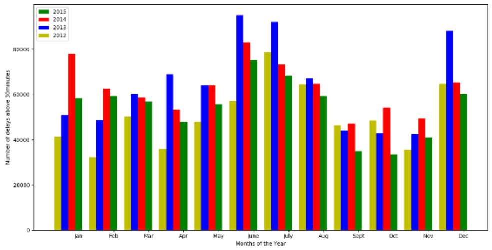


Figure 4: Month vs Delay for 2012, 2013 2014 and 2015

The reasons behind the above results was surprisingly straight-forward. June and July months were found to have more delays due to summer break and summer storms and delays in December accounted for snowfall and snow storms.

To compute the precise delay changes with respect to the month of the year, we performed a correlation analysis between datasets of year 2014 and 2015 and obtained promising correlation value of 92.3% which indicates better performance of our classifiers. The same experiment was performed on delays with respect to airports on 2014 and 2015 datasets and a correlation value of 87.5% was found

Figure 5 represents a boxplot between month and delays for the year 2015. From the box plot, it is clear that there are no outliers. The calculated IQR (Interquartile Range) is 11,480.

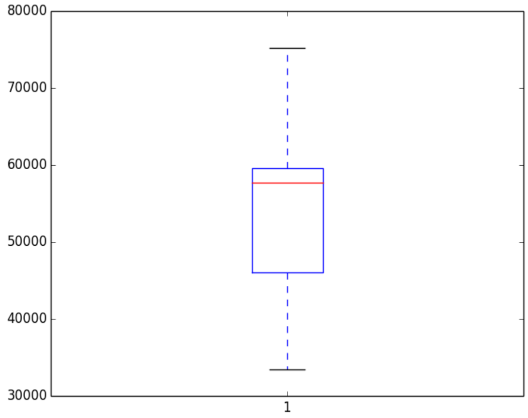


Figure 5: Boxplot - Number of delays vs month for the year 2015

Then we tried to plot a box plot between day of the week and delays for the year 2015. From Figure 6, we can figure out that there is one outlier. Saturday is possibly an outlier because it has the least number of delays in the week. The calculated IQR (Interquartile Range) is 11,075.

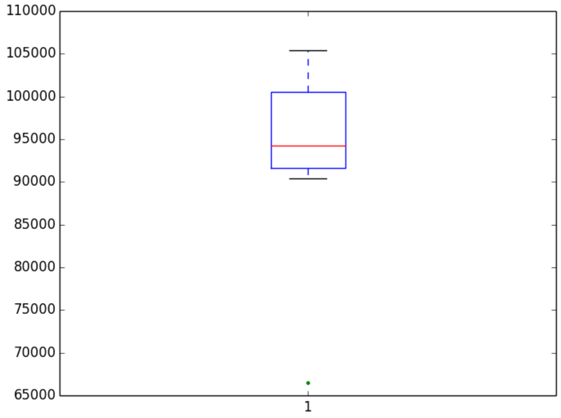
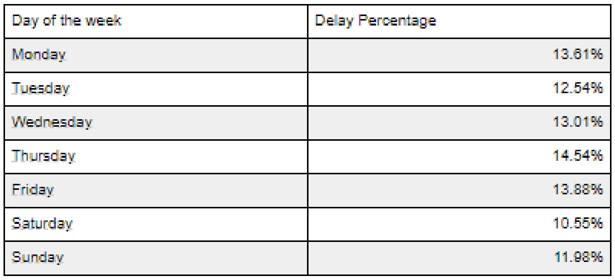
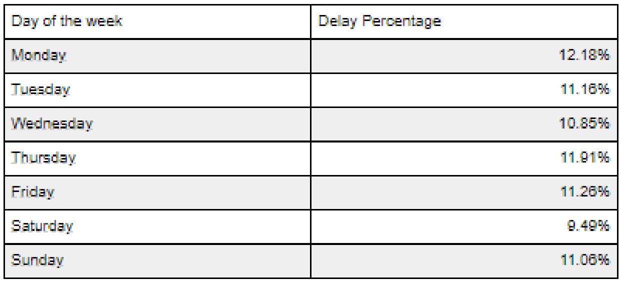


Figure 6: Boxplot - Number of delays vs day of the week for the year 2015



Day of the week	Delay Percentage
Monday	13.61%
Tuesday	12.54%
Wednesday	13.01%
Thursday	14.54%
Friday	13.88%
Saturday	10.55%
Sunday	11.98%

Figure 7: Delay analysis for day of the week for the year 2014



Day of the week	Delay Percentage
Monday	12.18%
Tuesday	11.16%
Wednesday	10.85%
Thursday	11.91%
Friday	11.26%
Saturday	9.49%
Sunday	11.06%

Figure 8: Delay analysis for day of the week for the year 2015

We also made some delay analysis on days of the week for two years – 2014 and 2015. It is clear that Saturdays have less number of delays compared to other days. The reason behind this could be accounted for the fact that the traffic and congestion is less on Saturday because people usually plan their travel either during start and end of any particular week.

3.2 Average delay propagation

To evaluate the impact of the propagation delay caused by the previous flight to the next connected/dependent flight, we analyzed the delay distribution at the ORD Chicago airport for the year 2015 in the month of June due to the highest delay for these as shown in the Figure 4 and Figure 2. The average daily delays of the flights arriving to the ORD airport, the average daily departure delays of the flights departing from the ORD airport, the average daily arrival delays at all different airports from the ORD airport are compared to find the correlation among them. The higher the correlation, the higher is the delay propagation.

Average Delay Parameter 1	Average Delay Parameter 2	Correlation coefficient
Arrival Delay with destination as ORD	Departure delay with Origin ORD	0.95678853
Arrival Delay with destination as ORD	Arrival delay with Origin ORD	0.96411223
Departure delay with Origin ORD	Arrival delay with Origin ORD	0.96399138

Table 1: Average delay propagation for ORD airport

The variations of the above-mentioned delays are very coherent resulting in high correlations as shown in Table 1 and Figure 5. This provides a strong causality relation that a delay caused by the current flight has a very strong impact of causing a delay in the connecting flight.

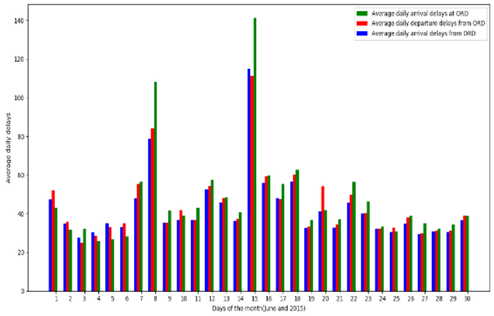


Figure 5: Average daily delays for June month showing delay propagation

3.2 Classifier analysis

We considered three different classifiers for our flight delay analysis – Naïve Bayes, Logistic Regression and Random tree classification. In all our experiments, we consider test data to be the flight dataset from the year 2015 and datasets from 2012, 2013 and 2014 were considered to be training data. In order to evaluate their performance, we considered accuracy to be an important factor. The delay threshold that we set was 30 minutes. So, any flight that is delayed for 30 minutes beyond the scheduled arrival time is considered delayed d_{min} when we train the classifiers. The accuracy for Naïve Bayes, Logistic regression and Random tree classifier is as shown in Table 1.

Classifier	Output Accuracy
Logistic Regression	0.80345
Random tree classifier	0.86420
Naïve Bayes	0.79725

Table 2: Accuracy of different classifiers on the 2015 data set with a d_{min} equal to 30 minutes.

From the table, we found Random tree classifier's accuracy to be the highest. Hence we chose Random tree classifier for all our future analysis.

3.2.1 Training-data subsets

We created subsets of training data to analyze the performance of Random tree classifier. Subset S_1 contained only year 2012 dataset, S_2 contained 2012 and 2013 datasets and S_3 contained 2012, 2013 and 2014 datasets. The accuracy of the classifier when we consider these three subsets as our training data is as shown in Table 2.

Train data	Test data	Accuracy
2012	2015	0.848457767337
2012,2013	2015	0.862914070745
2012,2013,2014	2015	0.864201392435

Table 3: Accuracy of Random tree classifier for subsets S_1, S_2 and S_3

3.2.2 Feature set

We selected a variety of features present in our dataset to see which provides us with a better accuracy. Accuracies for different feature sets is as shown in Table 3.

Training features	Accuracy
'DEST_AIRPORT_ID', 'MONTH', 'DAY_OF_MONTH', 'AIRLINE_ID', 'ORIGIN_AIRPORT_ID'	0.864201392435
'DEST_AIRPORT_ID', 'MONTH','DAY_OF_MONTH','ORIGIN_AIRPORT_ID'	0.872152997651
'DEST_AIRPORT_ID', 'DAY_OF_MONTH', 'ORIGIN_AIRPORT_ID'	0.872280665918
'DEST_AIRPORT_ID', 'MONTH', 'ORIGIN_AIRPORT_ID'	0.879066234297
'DEST_AIRPORT_ID', 'MONTH','ORIGIN_AIRPORT_ID','WEATHER_DELAY'	0.88585393048

Table 3: Accuracy of Random tree classifier for various feature sets.

From the table, we can see that when we include features with a set involving destination airport, month, origin airport ID and weather delay, we obtain the highest accuracy. This is probably due to the fact that delays happening due to weather has a major percentage to contribute towards overall delay.

3.2.3 Delay threshold evaluation

We set four threshold limits – 15 minutes, 30 minutes, 60 minutes and 90 minutes for flight delays. For our analysis, we considered origin-destination airports, month and weather delay as our feature set as this feature set gave us the highest accuracy as previously stated. Table 4 shows various accuracies for the above-mentioned thresholds.

Delay threshold (mins)	Output accuracy
15	0.816100245123
30	0.88585393048
60	0.9428152555067579
90	0.967719078746

Table 4: Accuracy for various threshold limits - $Accuracy_{15} < Accuracy_{30} < Accuracy_{60} < Accuracy_{90}$

From our analysis on the accuracy of threshold limits, we found out that higher delay threshold limits result in better accuracy of our classifier. In other words, we see a pattern where the likelihood of a given destination-origin airport having a flight delay of 90minutes is more for any given year. This analysis helps us to predict the occurrence of flight delays for higher threshold.

The regression model we develop is used to predict various flight delays based on factors such as the time of the year, airport traffic, flight and weather. Since we are testing our model on the 2015 dataset, we propose to evaluate our model based on its accuracy to predict the delays that occurred in the year 2015. Keeping all the other factors constant, we can also evaluate our model on the flights scheduled to occur on the present day. Additionally, since we are also suggesting alternate routes, we should also ensure that the journey time for alternate route should not exceed by a considerable margin, the journey time of the delayed flight plus the delay itself.

4. EVALUATION AND RESULTS

Initially, we had several brainstorming sessions to narrow down our project idea within the team. The travelling domain was one of the common interests among all the team members and hence we finalized about going with flight delay prediction. The project presentation with Professor Qin LV was very useful. We were restricting our dataset with 2008 and 2015-year flight delay data from Kaggle. Based on the discussion with Professor, one year's data would give unpredictable results due to specific events happening in that year. This made us look for more historical data from the Bureau of Transportation Statistics. Professor also mentioned about the cost while suggesting for alternative routes. Based on this, we are trying to find a dataset which would give details about the cost along with the flight details.

5. Conclusion

We have performed a thorough analysis on the flight delays and its correlation with multiple factors such as airports, month of the year and the airlines. During the analysis, significant flight delays were seen in the month of June, Chicago airport, EV airlines independently. The results were promising with respect to the practical conditions such as summer storms, congestion, airport traffic, region of airline operation.

The next significant analysis was done by building flight delay occurrence prediction model through various classifiers such as Naïve Bayes, Decision tree and Logistic regression. The classifiers and the features were compared and the decision tree classifier turned out to have the highest accuracy of 86%. The best features to train the model included origin airport, arrival airport, month, weather delay. By building multiple classifiers for 15min, 30mins and 60mins we were able to mimic the linear regression model with discrete values.

The delay propagation is a crucial analysis in this paper as many flights are delayed due to the delay in the previous connected flights. The above analysis was done on the ORD airport for the June month with the highest delays, the result in the correlation coefficients between the average daily delay arriving to the ORD, average daily departure delay from ORD and the average daily arrival delay from ORD were very significant and promising with correlation values more than 95%.

6. FUTURE WORK

In the future, we are planning to include additional weather data as features for classification. Additional weather data includes details such as snow, precipitation level, wind speed, wind direction etc. We are also planning to create a graph with the data set and predict better alternative routes to the user. We are also planning to build a mobile application or a web application so that the user can input his flight details and get information about flight delay. For doing this the dataset should be placed in a cloud like AWS and a web service should get the input and output the delay information and alternative flight routes. And also, we are planning to scale the application using different technologies like Spark. By setting up a spark cluster we can split the workload across multiple nodes which will improve system's performance.

7. REFERENCES

[1] Juan Jose Rebollo and Hamsa Balakrishnan *"*Characterization and Prediction of Air Traffic Delays*"*

[2] Young Jin Kim, Sun Choi, Simon Briceno, Dimitri Mavris "A deep learning approach to flight delay prediction", 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)

[3] Schaefer, L. and D. Millner, October 2001, "Flight Delay Propagation Analysis with the Detailed Policy Assessment Tool," Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference, Tucson, Arizona, Institute of Electrical and Electronics Engineers (IEEE).

[4] Abdelghany, K. F., Abdelghany, A. F., and Raina S., (2004) "A model for projecting flight delays during irregular operation conditions, Journal of Air Transport Management", Volume 10, Issue 6, Pages 385-394

[5] Allan, S.S., S.G. Gaddy, and J.E. Evans, (2001)" Delay Causality and Reduction at the New York City Airports Using Terminal Weather Information", MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Lexington, Massachusetts

[6] NOAA – National Oceanic and Atmospheric Administration for Weather dataset

[7] The Bureau of Transportation Statistics for flight delay dataset

8. Appendix

Individual contributions -

Rakesh Margoor: Requirement analysis, Data collection, Classification and delay analysis, Bar graphs, correlation analysis, data preprocessing, report writing

Avinash Ratnavel: Requirement analysis Code related to statistics – Boxplot, Bar graphs analysis of classifiers, Data cleaning, report writing.

Chirag Kamat: Requirement analysis, Delay threshold analysis, design, delay propagation analysis, classifier selection, data preprocessing, report writing

On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.
Avinash Ratnavel Maharaj, Chirag Kamat, Rakesh Margoor