Rama Krishna Valluri 1001278967

Ramabhadra Raju Tirumalaraju 1001268358

Viral Mehta 1001154251

Ramasundari Baalakrishnan 1001276206

May 2, 2016

PREDICTING FLIGHT DELAYS

Table of Contents

[**Articulation of the problem** 2](#_Toc449958536)

[Objective 3](#_Toc449958537)

[**Data collection and preparation** 3](#_Toc449958538)

[Feature selection 4](#_Toc449958539)

[Column transformation 4](#_Toc449958540)

[Target 5](#_Toc449958541)

[**Analysis** 6](#_Toc449958542)

[Exploratory analysis 6](#_Toc449958543)

[Predictive Analysis 9](#_Toc449958544)

[Cost Analysis 12](#_Toc449958545)

[Twitter Analysis 13](#_Toc449958546)

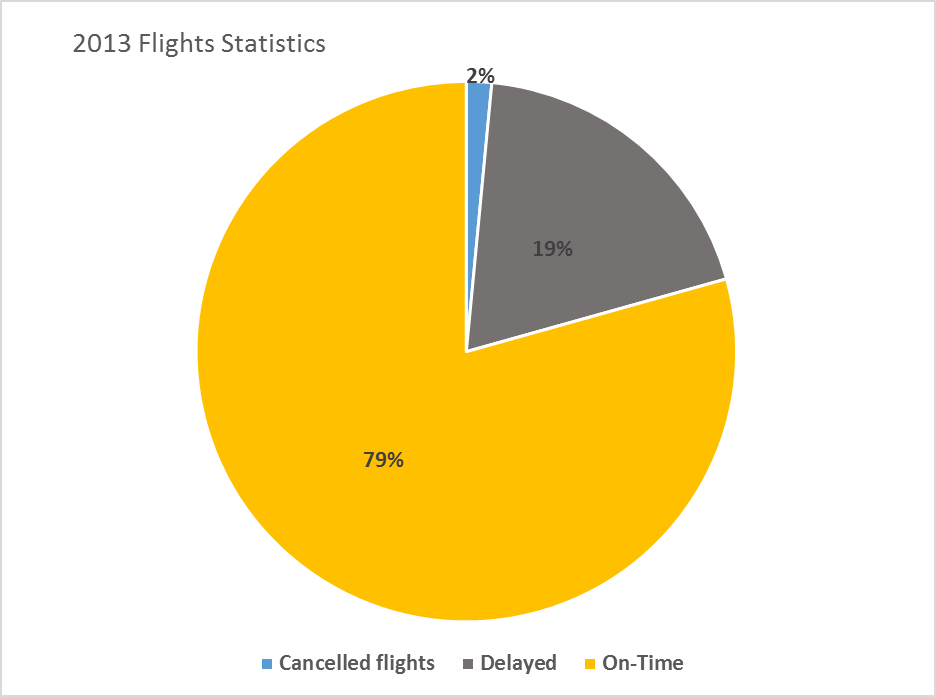
[**Limitations** 13](#_Toc449958547)

[**Conclusions** 14](#_Toc449958548)

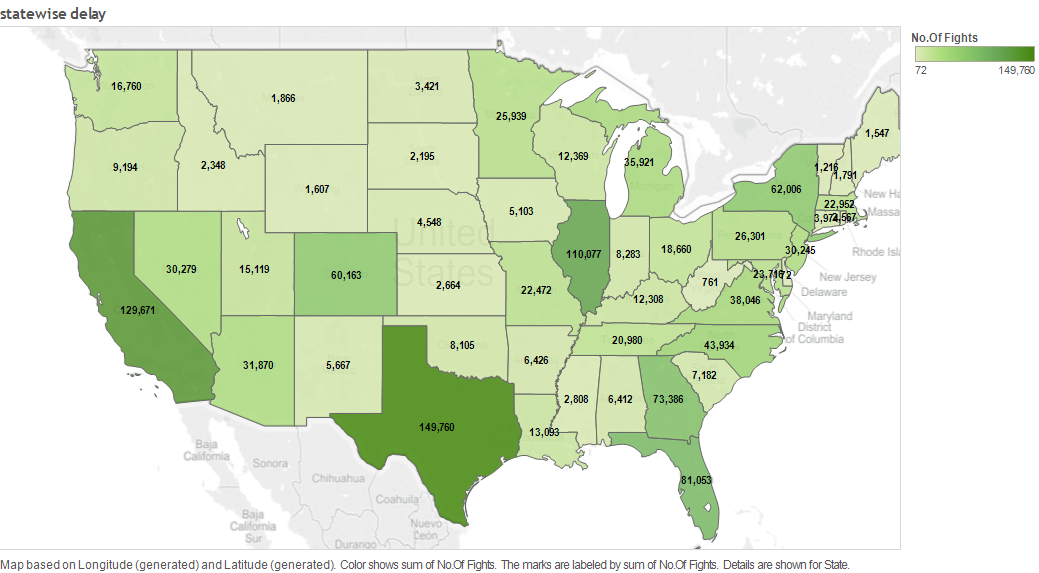
**PREDICTING FLIGHT DELAYS**

# **Articulation of the problem**

Out of 8 million commercial flights that operate in US each year, 25% of them are delayed. The Federal Aviation Administration assessments show that these flight delays cost airlines $22 billion every year. The Aviation authorities and aviation need to know the factors leading to specific flight delays so that they can take the required actions to ensure better operations. Also, it supports for a better planning if the passengers know about any kind of disruption in their journey. The below pie chart shows the flight delays for the year 2013. Out of 6,369,482 flights 1,219,890 were delayed.



As per the state wise flight delay data, Texas has the highest flight delays for the year 2013 (149,760). Out of which the maximum delays were in DFW area, so we have considered DFW for our predictive analysis. Apart from providing good prediction for the entire model, this also, reduces the dataset.



## Objective

Objective of the project is predicting delayed arrival of flights for Dallas Fort worth airport.

Arrival delays of flights are defined as the flights that arrive 15 or more minutes late.

# **Data collection and preparation**

Source links for the dataset were mentioned by Crowdanalytix, the competition host website.

The base dataset consists of on-time performance data (historical flight data) for the US domestic flights and the corresponding weather data.

The size of the entire dataset was 28 GB. This was loaded into **Microsoft SQL Server Management Studio**. Data specific to DFW was extracted using SQL queries. (Please find the scripts in the SQL Query File).

Dataset included facts for 347 airports and 20 carriers. For DFW airport, there were around 550,000 records for flights with origin and destination as DFW airport. To build the logic of our model we worked on 25,000 records randomly selected from the above half a million records for DFW airport.

Historical data consists of 42 parameters. It has features on-time performance flight details like month, unique\_carrier, origin\_city\_name, dest\_city\_name, arr\_del15, arr\_delay\_group, etc.

We also had data related to weather, traffic data, stock conference holidays which we have included as described below.

## Feature selection

Selection of the features from historical data for analysis:

We selected the attributes that were most important for our analysis. Following is the list of features that were included for our analysis:

|  |  |
| --- | --- |
| month | Month |
| day\_of\_month | Day of Month |
| unique\_carrier | Unique carrier code |
| dest\_city\_name | Destination Airport, City Name |
| crs\_dep\_time | CRS Departure Time Block, Hourly Intervals |
| crs\_arr\_time | CRS Arrival Time Block, Hourly Intervals |
| crs\_elapsed\_time | Scheduled Duration |
| Delay | Delay in minutes |

We have excluded the following features as this information will not be available for the flights that have not yet departed:

crs\_arr\_time, arr\_time, arr\_delay\_new, arr\_del15, arr\_delay\_group, arr\_time\_blk

## Column transformation

1. day\_of\_month: This column described the day of month which were numeric values. This was an important feature for our analysis. We have the stock conference holidays dataset that provides the information for holidays.

Based on the holiday data we converted the day\_of\_month into a categorical column with values as

WEEKEND: Friday, Saturday, Sunday.

PreHoliday: Two days before and after the holiday.

Weekday: The remaining days of the week.

1. crs\_dep\_time: This column provided the departure time of flight in the form of HHMM. This was a numeric attribute. We converted the values into categories of hours. For eg. A record with 1530 value was converted to 3 since it belonged to the 3 p.m block.

The new column was named as ‘dephrs’. Similarly crs\_arr\_time was transformed to ‘arrhrs’.

1. crs\_elapsed\_time was named to ‘schduration’. Apart from the schduration, all the columns are categorical.

## Target

Our dataset consisted of multilabel target as follows:

carrier\_delay, weather\_delay, nas\_delay, security\_delay, late\_aircraft\_delay

We have combined these columns into a single class column called ‘delay’ which would have values ‘No Delay’ if the flight was delayed less than 15 mins and ‘Delay’ if the delay time was more than or equal to 15 mins.

The final historical dataset consisted of 25,000 records for 2013 year for DFW Airport with the features as below:

month, WorkDay, unique\_carrier, dest\_city\_name, dephrs, arrhrs, schduration, delay.

**Including Weather Data:**

The total size of weather data was about 4.41 GB.It consisted of information related to weather specific to Airports, carriers and flights.

We selected the following features to be important and considered for adding to our final historical dataset:

temperature\_f (Temperature in Fahrenheit), visibility\_mph, conditions.

Handling of weather data was done using MS SQL and script was written to merge the weather data with the historical data.

While adding weather to historical data, missing values were found in the combined files. We excluded the missing value records.

The final historical dataset with weather data consisted of 25,000 records for 2013 year for DFW Airport with the features as below:

month, WorkDay, unique\_carrier, dest\_city\_name, dephrs, arrhrs, schduration, temperature\_f, visibility\_mph, conditions, delay.

At first, the historical dataset was used for predicting delay of the flights. Weather data was then included to the historical data to check if there was an improvement in prediction.

# **Analysis**

## Exploratory analysis

**Average arrival delay by month:**

When the impact of month on flight delays is considered, it would be expected that holiday seasons would have more delays due to the traffic during that period. Bar graphs with average delay in minutes versus month is plotted for arrivals and departures.





For both arrivals and departures, we can observe more delays during the month of June and July; the summer holidays period where the number of flights are more. The delays during summer are owed to the thunderstorms in DFW area, unlike winter storms summer storms pop up suddenly and can’t be predicted way before [1]. Also, the impact of December month is visible. The delays during the period March, Sept, Oct, Nov is the least.

Ref: [1] <http://articles.sun-sentinel.com/2011-06-19/news/fl-summer-flight-delays-20110619_1_airline-delays-winter-storms-thunderstorms-move>

**Delays by time of the day:**

The hour of the day also has an effect on flight delays, a bar graph for average delay versus the hour of the day is plotted.





From the plots we can clearly observe that as the hour of the day proceeds the flight delays increase. This is due to the accumulated delays of flights from the morning; the delay of one flight affects the next flight and so on. From the plots the peak average delay for departure flights is at 20:00 and for arrival flights is around 22:00. As the delay increases along with the hour of the day this can be a good predictor of the model.

**Delay by carrier:**

There is possibility that Airline carrier can also cause some effect on the delays. A bar graph for average-delay versus carrier for departures and arrivals is plotted below.





From the plots we can see that small and low cost carriers such as Atlantic south east airlines(EV) JetBlue(B6) are delayed more. On the other hand, main stream carriers such as Alaska airlines(AS), delta(DL) have smaller delays.

## Predictive Analysis

For predictive analysis the dataset was split into two parts: Arrivals and Departure from DFW. As mentioned earlier this was done to find the effect of city from/to which the aircraft has arrived/departed. The initial analysis was done on the departures from DFW and a random sample of 25k records were chosen from the departure data (270k)for computational easing.

The following machine learning algorithms were used in our analysis:

1. Logistic regression
2. Decision Tree
3. Random Forest

We have used dummy variable encoding as all the variables chosen were categorical except scheduled duration of the flight. There was an improvement of 4% accuracy across all the algorithms. Also, cross validation with stratified K=5 fold was used.

As the dataset was having imbalance in the target class, SMOTE technique was used to oversample the minority class (Delayed). Also it is more important to predict a delayed flight.

**With Historical flight data**

|  |  |  |
| --- | --- | --- |
|  | **Accuracy(%)** | **recall delayed (%)** |
| **logistic** | **64.44** | **0.35** |
| **decision tree** | **68.09** | **0.47** |
| **random forest** | **72.36** | **0.57** |

Please find the outputs of the algorithms in results.docx file

The below is the feature importance result from logistic regression

|  |  |
| --- | --- |
|  | Coefficients |
| dest\_city\_name\_Newark | 0.84 |
| unique\_carrier\_EV | 0.74 |
| Month\_7 | 0.37 |
| Dephrs\_19 | 0.36 |
| Arrhrs\_20 | 0.31 |

Looking at the feature selection more flights are likely to delayed in summer (month\_7/July), evening flights are more delayed. More analysis on the feature importance will be done once the weather data is added.

**With Weather data included**

|  |  |  |
| --- | --- | --- |
|  | **Accuracy (%)** | **Recall delayed (%)** |
| **logistic** | **66.12** | **0.40** |
| **decision tree** | **70.53** | **0.55** |
| **random forest** | **74.78** | **0.57** |

Please find the outputs of the algorithms in results.docx file

As we can see, after adding the weather data there was improvement of about 2% in accuracy of all the models. The recall for delay also improved logistic and decision tree. Random forest had no improvement in the recall for delay, but its accuracy improved by 2% which meant that it is predicting more no delay flights correctly.

The below is the feature importance result from logistic regression

|  |  |
| --- | --- |
|  | Coefficients |
| dest\_city\_name\_Honolulu | 0.99 |
| unique\_carrier\_AA | 0.95 |
| Dephrs\_23 | 0.95 |
| Arrhrs\_0 | 0.88 |
| Month\_6 | 0.78 |
| conditions\_'Light Thunderstorms and Rain' | 0.69 |

From the feature importance table we can relate to the analysis done in data exploration. The model was taking into account the destination city, carrier, arrival/departure hours and conditions. As figured out in the exploratory analysis flights arriving or departing in the late hours had more delays, month\_6 (june-summer) had more delays. Carrier wise American airlines seems to have more delays. Regarding conditions light thunderstorms is causing more delay than Heavy thunderstorms this is not intuitive, but it may be due to the fact that heavy storms can be tracked beforehand and departures/arrivals can be adjusted accordingly.

In Random forest the order of importance was: temperature, schduration, visibility, WorkDay, conditions, month. As we analyze this order we can see that the model is actually taking into account more weather related factors, this explains the high accuracy of random forest model.

By the analysis of results random forest gave the best accuracy and the recall for delays was also high, so we chose this model. Below are the results for random forest with weather data.

**Accuracy**: 0.7478

**The confusion matrix is**:

No Delay [[16274 2697]

Dealy [ 5127 6931]]

**Cost analysis** (1-No Delay 5-Delay) is: 28332

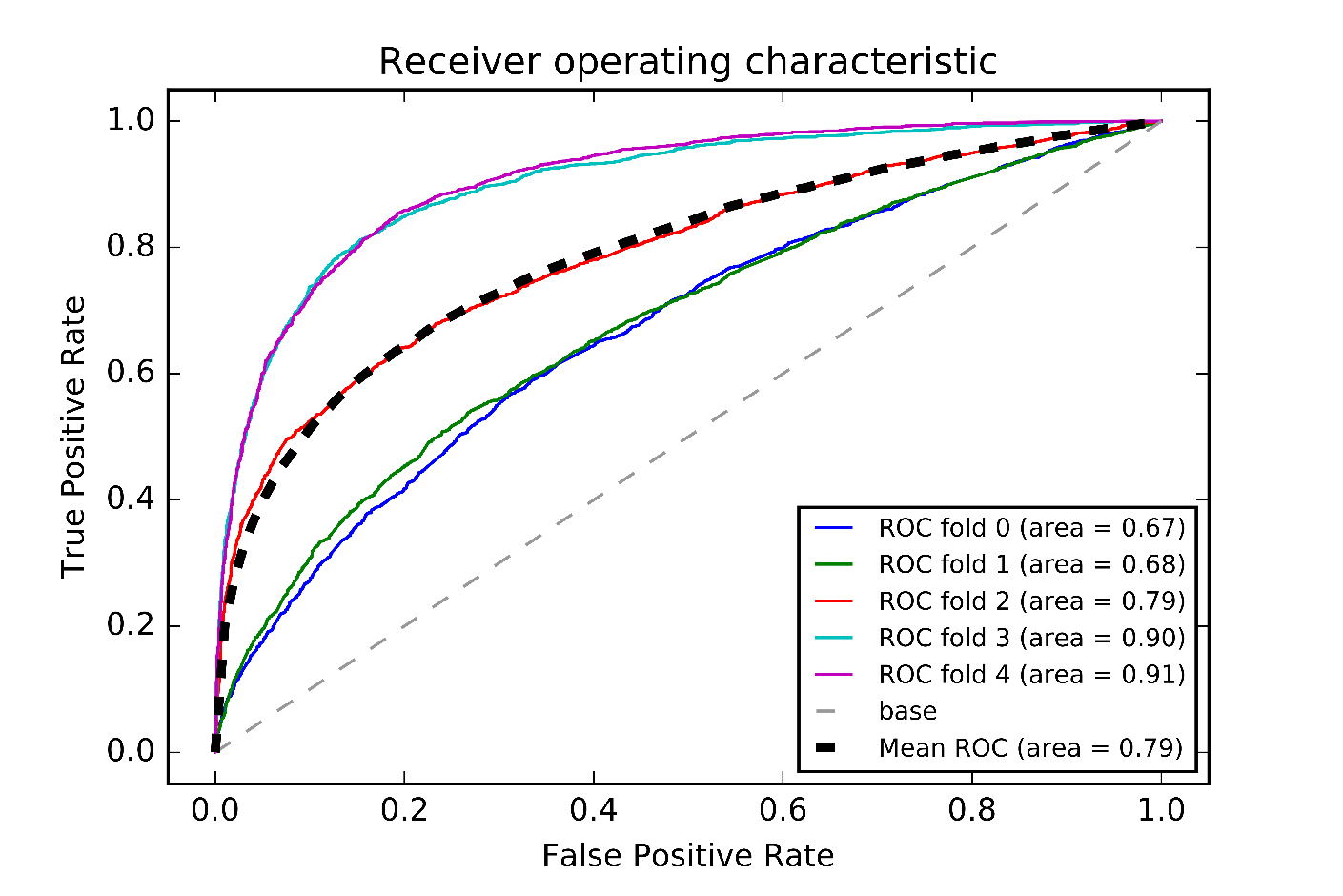
**The classification report is**:

precision recall f1-score support

No Delay 0.76 0.86 0.81 18971

Delay 0.72 0.57 0.64 12058

avg / total 0.74 0.75 0.74 31029



Looking at the ROC curve for random forest we can see that the variance is high when compared to logistic regression or decision tree (Appendix). This can be attributed to the complexity of random forest model, choosing a voting mechanism from different trees. Less complex model logistic regression had the lowest variance among all.

In the best model that we have if we do a cost analysis based on the reasoning that classifying a delayed flight as on-time leads to 5 times more cost compared to classifying an on-time flight as delayed.

## Cost Analysis

The goal is to predict whether a flight has been delayed or not.

We believe that the cost of predicting a flight as ‘No Delay’ when it is actually delayed, is much higher than predicting a flight as ‘Delay’ when it is not.

We do the cost analysis based on the confusion matrix that provide the information of wrongly predicted values for ‘Delay’ and ‘No Delay’0 here we see that 1421 records were falsely predicted as No Delay.

In general the average cost per flight delay is about $ 11300.

Cost Matrix:

|  |  |  |
| --- | --- | --- |
| Delay | 0 | 1 |
| No Delay | 5 | 0 |

Considering the above cost matrix we assume that the cost of wrongly predicting a flight as ‘No Delay’ is 5 times that of wrongly predicting a flight as ‘Delay’. So the cost of predicting a single flight as No Delay is about $2260. We tried to adjust the class weight to get the least cost from our confusion matrix.

The confusion matrix is:

|  |  |  |
| --- | --- | --- |
|  | No Delay | Delay |
| True | 6460 | 12511 |
| False | 1421 | 10637 |

We get the least cost for the above confusion matrix which is about 19616 and the accuracy is about 55%.

## Twitter Analysis

Tweets were collected using Hashtags as #DFW delays or #DFW delay for a single day (26th April 2016). These HashTags were specified to get the tweets related to the delays that occurred at DFW Airport. Tweet sentiment was calculated using TextBlob in python and classified as negative tweet if the sentiment polarity is less than 0.

Flight delay data was collected for 26th April and analyzed against the negative tweets.

The hour’s column indicate the time of the day. For eg. 15 indicates 3:00 pm. We observe the sentiment score line against the no of flights delayed. For hours where the no of flights are delayed more, the sentiment score of tweet is more negative. Hour 21 (9:00 pm) has one of high fight delays and the twitter sentiment is most negative during that period.

Please find the related files in the TwitterAnalysis folder.

# **Limitations**

* The collected dataset contains only origin and destination, the in transit is not considered, which might contribute to the delay. The model accuracy might increase if this is taken into account.
* Also, the delay type in the dataset is sub-categorized into aircraft arriving late delay, national aviation system delay, air carrier delay, weather delay, security delay. We couldn’t consider all delays, as each flight is prone to delay from one or many factors. However, this can be considered if we build a model using multi-label problem.
* Attributes such as traffic; both traffic by carrier and traffic by originating and destination airports, stock price of airline carriers and tail number of flights were not taken into account due to the enormity of data. However, considering these factors we hope the model can be predicted better.
* Attributes such as bird strike account to unexpected delays, these should also be taken into consideration based on the seasonality and bird count near airports.

# **Conclusions**

As per the analysis done, attributes such as month, hour of the day, carrier and weather plays an important role in predicting delays. Months during summer and December; holiday season were more likely for a delayed flight. Unpredictable nature of thunderstorms during this period also accounts for the delay.

Also, Delays during the peak hours; 18:00-22:00 were common. Accumulated delays from morning flights augments to flights in the later hour of the day, this creates a chain of delays and can become unmanageable. So airline authorities should make required amendments such as better management during holiday season, hiring additional staff to reduce the effect of preceding flights on subsequent ones.

As per the analysis, low cost airlines account for much delay when compared to established carriers. The low cost airlines, apart from providing economical flights, should also try to provide better services to sustain in competitive aviation industry.

The best model is predicted by random forest after considering the weather data, so weather plays a significant role in predicting the delays. Comparatively, the delays due to light thunderstorms were more than heavy thunderstorms, this is due to the unpredictable nature of light thunderstorms which rise abruptly and mostly unnoticed prior hand. Better prediction of light thunder storms will help the aviation industry in decreasing the delay of flights.