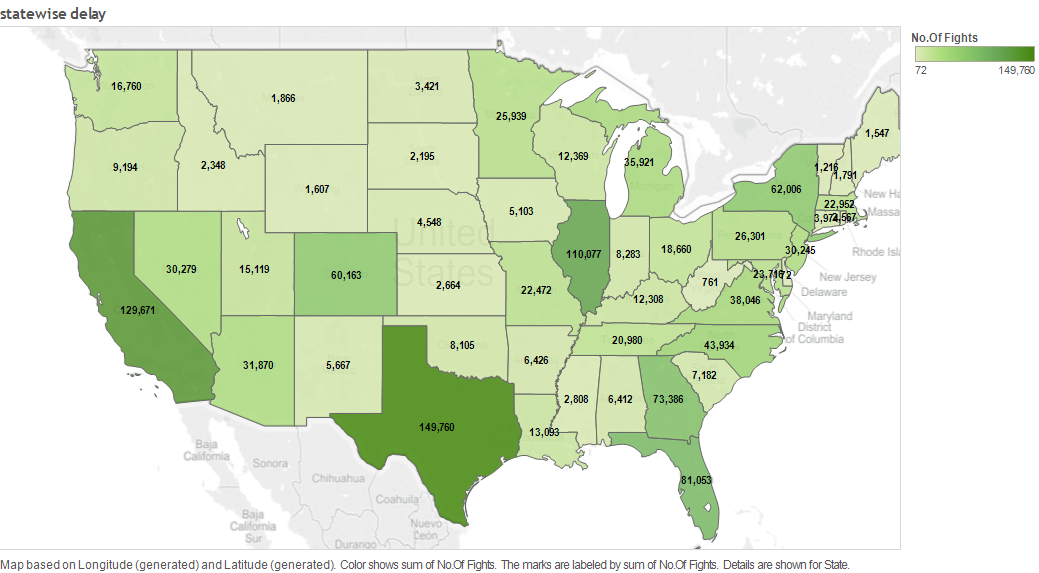
**DATA SCIENCE PROJECT REPORT**

**PREDICTING FLIGHT DELAYS**

**a) 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.

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 a flight.

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

**b) Data collection and preparation**

Source links for the dataset were mentioned by Crowdanalytix, the competition host website. There were no missing values.

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 refer to the SQL scripts in the Index page).

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 24,000 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.

Other data- weather data inventory data, traffic data, stock conference holidays.

**Feature selection:**

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 |
| origin\_city\_name | Origin Airport, City Name |
| dest\_city\_name | Destination Airport, City Name |
| Dephrs | CRS Departure Time Block, Hourly Intervals |
| Arrhrs | CRS Arrival Time Block, Hourly Intervals |
| Schduration | Scheduled Duration |
| Distance | Distance travelled |
| 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:

.,,/.,,/.,.,/

Training dataset: 2013 historical data for DFW airport.

Test dataset: 2014 historical data for DFW airport.

**Data compilation:**

At first, the above training and test datasets were used for predicting delay of the flights. Then weather data was included to the historical data to check if there was an improvement in prediction.

Data split: While running the prediction algorithm we do it once for the arrival and once for departure data for DFW. This is done as we cannot see any interaction between arrival and departures. When we included weather data there were missing values for temperature and visibility.

**Target class:**

1. delay type(Yes or No)

**c) Analysis and results**

**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** |

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** |

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 maybe 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

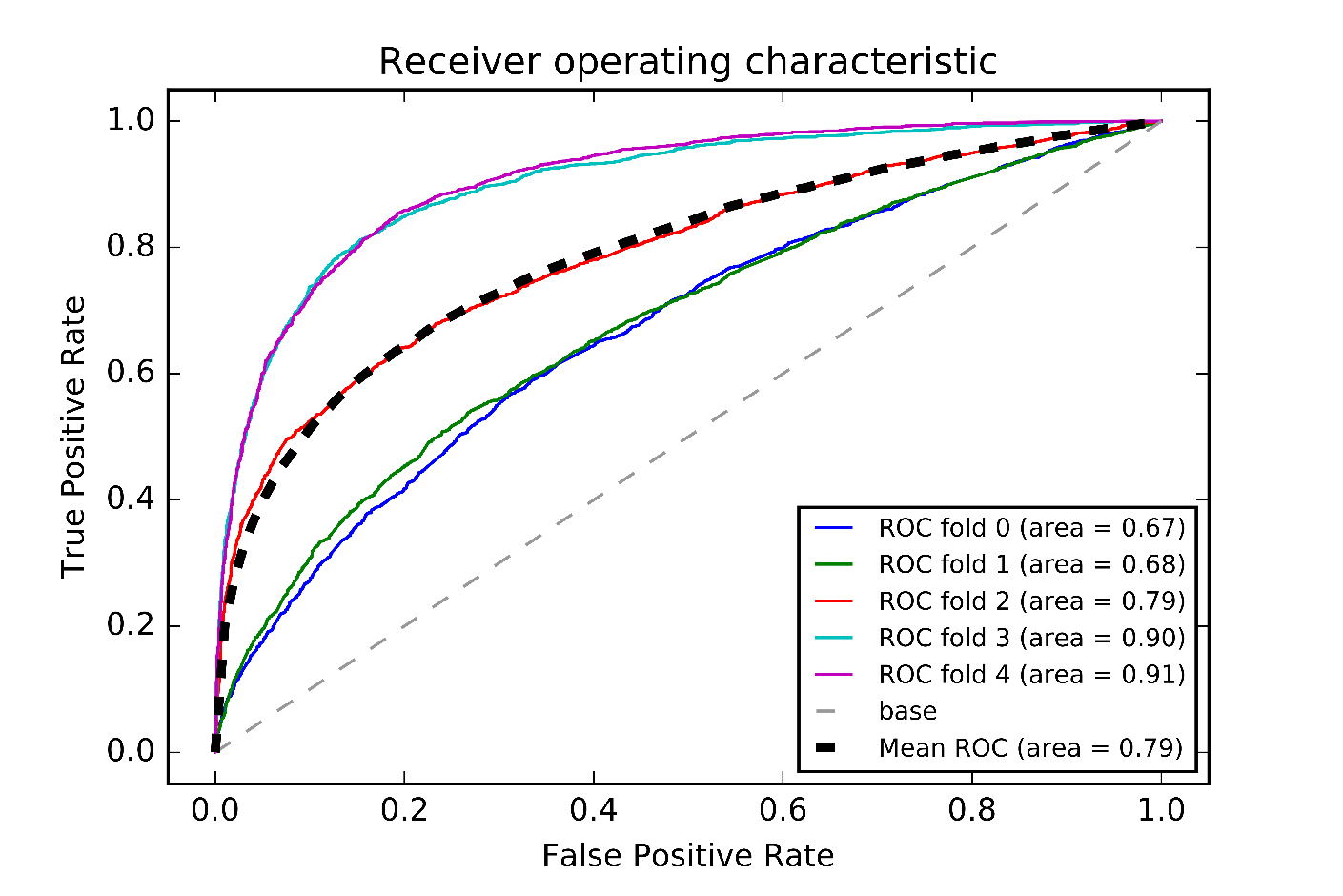
**The classification report is**:

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

The cost is high in this model and in the next section we will do a cost analysis to reduce it.

**Twitter Analysis:**

**d) Limitations**

Prediction model considers an origin and one destination. (No transits/ legs)

The delay type can be categorized based on if it is an aircraft arriving late delay, national aviation system delay, air carrier delay, weather delay, security delay.

We did not predict the delay type as a delay for a flight can have more than one type of delay and we had to use Multi-label option.

**e) Conclusions/Implications/Directions for the future**