DATA SCIENCE FINAL PROJECT

Prediction of Flight Delays

TEAM MEMBERS:

STUDENT NAME	STUDENT ID
Bhargav	200207466
Kalyan Ghosh	200201466

Contents

1 INTRODUCTION

Delay can be considered as the most important feature to measure the performance of any transportation system. Delay can be defined as the difference between the scheduled departure time and the actual departure time (Actual Dep-Schedule Dep).

The aviation industry forms an important component of modern urban and semi urban transportation system transporting people from one place to another in the fastest time possible as compared to other transportation systems like roads and railways. But the aviation industry also suffers from flight delays resulting out of numerous factors like bad weather, traffic etc. In 2013, 36% of flights were delayed by more than five minutes in Europe and 31.1% of flights were delayed by more than 15 minutes in the United States. Flight delays often lead to negative impacts, mainly economic for passengers, airlines and airports. Due to this uncertainty, passengers usually plan to travel many hours before their appointment, increasing trip time and costs. On the other hand, airlines suffer penalties, and increased operational costs. Furthermore, from the sustainability point of view, delays may also cause environmental damage by increasing the fuel consumption and gas emission. Hence accurate prediction about the delay of flights can bring much respite to flyers who can plan their journey more efficiently and it can also help the airline carriers by reducing their operational costs.

In this project, we present an analysis of the flight delay prediction from a Data Science perspective. In our project we use Random Forests classifier as the Machine Learning method to model flight delay predictions. We specifically have chosen Random Forests because of ability of balancing error in class population unbalanced datasets which is important in our project because the dataset that we considered in our project is highly imbalanced with the number of delayed flights only forming a small fraction of the entire dataset.

2. LITERATURE SURVEY:

From the literature we reviewed, we found that this problem of flight delay prediction has been approached differently by different people. The flight delay prediction problem may be modelled by many methods, depending on the objectives of the researches. The methods are divided into 5 groups: Statistical Analysis, Probabilistic Models, Network Representation, Operational Research and Machine Learning. In this project we approached the problem from machine learning perspective. Researches that study flight systems from machine learning perspective are increasing. The methods commonly used include k-Nearest Neighbour, neural networks, SVM, fuzzy logic, and random forests. They were mainly used for classification and prediction.

Rebollo et al. [1] applied random forests to predict delay innovation. They compared their approach with regression models to predict delay innovation in airports of the United States considering time horizons of 2, 4, 6 and 24 hours. Lu et al. [2] compared the performance of Naïve Bayes, decision tree and Neural networks to predict delays in large datasets. They observed that decision trees had the best performance with confidence of 80%.

Lu et al. [3] built a recommendation system to forecast delays at some airports due to propagation effects. Prediction was based on the k-Nearest Neighbour algorithm and used historical data to recognize similar situations in the past. The authors noticed fast response time and easy logical comprehension as the main advantages of their method. Khanmohammadi et al. [4] created an adaptive network based on fuzzy inference system to predict delay innovations. The predictions were used as an input for a fuzzy decision making method to sequence arrivals at JFK International Airport in New York.

3.DATA:

For our project, we used the publicly available Flight dataset from the Bureau of Transportation Statistics(BTS) website[Dataset].

The website has more that 25 years of flight data with over 50 flight features. In our project, we consider the data from the year 2013 to 2016 and the size of our dataset is ~100k. From the numerous features we have identified the 9 most important features which we think has a substantial effect on the delay of a flight. The 9 input features in the dataset are as follows:

- Year: The year of the flight data
- Month: The month of the flight data
- DayOfMonth: The day of the month of flight data
- DayOfWeek: The day of the week of flight data
- DepTime: The departure time in 24Hrs format
- UniqueCarrier: The Airline Carrier
- Origin: The Origin city airport of the flight
- Dest: The Destination city airport of the flight
- Distance: The distance in miles between the Origin & Destination
- Dep_Delayed_15: This is the output feature of the data. It contains binary values ‘Y’ and ‘N’. This field is an indication whether a flight is delayed or not given the 9 input features.

A snapshot of the input dataset is shown in the below figure:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Year	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance	dep_delayed_15min										
2	2015	c-11	c-22	c-3	1316	AA	FAT	DFW	1313	N										
3	2013	c-10	c-12	c-2	1242	US	SAN	PHL	2369	N										
4	2015	c-5	c-5	c-5	1302	OO	SBP	SFO	191	Y										
5	2014	c-8	c-8	c-1	1138	DL	PHL	ATL	665	N										
6	2014	c-4	c-3	c-7	1438	DL	ATL	BDL	859	N										
7	2014	c-9	c-15	c-4	2311	OH	ATL	EVV	350	N										
8	2014	c-6	c-20	c-1	1208	NW	DTW	GRR	120	N										
9	2016	c-7	c-8	c-7	1839	WN	PDX	SMF	479	Y										
10	2014	c-2	c-27	c-7	1230	OH	PBI	CVG	892	Y										
11	2014	c-8	c-17	c-3	1723	UA	LAX	SFO	337	N										
12	2016	c-7	c-7	c-1	2115	DL	SLC	PDX	630	N										
13	2016	c-4	c-29	c-7	757	AA	EWR	SJU	1608	N										
14	2015	c-12	c-11	c-1	1838	XE	EWR	CLT	529	Y										
15	2014	c-10	c-3	c-1	1528	NW	RAP	MSP	490	N										
16	2014	c-11	c-22	c-2	713	US	ROC	PHL	257	N										
17	2014	c-12	c-26	c-1	926	NW	MDT	DTW	370	N										
18	2013	c-7	c-17	c-6	1736	NW	DTW	BWI	408	Y										
19	2014	c-2	c-14	c-1	1412	DL	BNA	ATL	214	N										
20	2015	c-7	c-11	c-2	1834	UA	DEN	FLL	1703	N										
21	2015	c-10	c-31	c-2	1143	UA	DCA	ORD	612	N										
22	2013	c-8	c-25	c-3	2059	HP	LAX	LAS	236	N										
23	2016	c-3	c-27	c-2	556	9E	PWM	DTW	668	N										

4.METHODOLOGY:

The high-level methodology used in this project is as follows:

- Initially, we downloaded the flight dataset from the Bureau of Transportation Statistics(BTS) website[Dataset]. For our project, we used the data from the year 2013 – 2016 and basic clean-up of data was done
- Exploratory analysis of the data was done to gain some insight into the features which are helpful in predicting the data.
- Data is prepared for the classifier by converting categorical features into dummy variable using one hot encoding.
- The dataset is then split into Training set and Test set in the ratio of 70% to 30% respectively.
- The Training set was used to perform a 3-fold Cross Validation to find the optimal set of hyperparamets for the Random Forest classifier
- We then, refit the Random Forest classifier to the entire Training set using the hyperparameter values obtained in the above step.
- Then the properties of the fitted classifier are evaluated on the Test dataset.

5. DATA PREPROCESSING:

Data was read into a *Pandas* data frame and basic data preprocessing was done as below:

- "Month"," DayofMonth" and "DayofWeek" fields were converted from string to integer data type.
- Records for which the departure time values were greater than 2400 were mapped back to the interval 0 to 2400.
- Records for which the "Year" values were not in the range of 2013-2016 were removed.
- Similarly, records for which "Months" were not between 1 to 12 were removed.
- Rows for which the "DayofWeek" values were not between 1 to 7 were removed.
- Rows which had invalid delay flag (Neither 'Y' or 'N') were also removed.

After the data preprocessing each record contains the following fields.

SERIAL NO:	FEATURES	DESCRIPTION	TYPE
1	Year	A number between 2013 and 2016	integer
2	Month	A number between 1 and 12	integer
3	DayofMonth	A number between 1 and 31	integer
4	DayOfWeek	A number between 1 and 7	integer
5	DepTime	A number between 0 and 2400	integer
6	UniqueCarrier	Two-character airline code	categorical

7	Origin	Three-letter departure airport code	categorical
8	Dest	Three-letter destination airport code	categorical
9	Distance	Flight distance in miles	integer
10	dep_delayed_15min	Y/N flag indicating a delay of ≥15 min	binary

5.1 CONVERSION OF CATEGORICAL VALUES:

The data for the Random Forest classifier is prepared by converting the categorical feature values ("UniqueCarrier"," Origin" and "Destination") into dummy values using the **one-hot encoding** technique.

5.2 CALULATION OF INTRINSIC DISCREPANCIES:

Then, we tried to summarize the effectiveness of each feature in its ability to distinguish between delay and non- delay. We do this by calculating the intrinsic discrepancy between the two probability distributions p1 and p2 for the different features. The formula that we use to calculate the intrinsic discrepancy as shown below.

$$\delta\{p_1, p_2\} = \min\left\{ \int p_1(x) \log \frac{p_1(x)}{p_2(x)}\, dx,\ \int p_2(x) \log \frac{p_2(x)}{p_1(x)}\, dx \right\}.$$

We then sort the top 10 intrinsic discrepancies in descending order and summarize them in the table below.

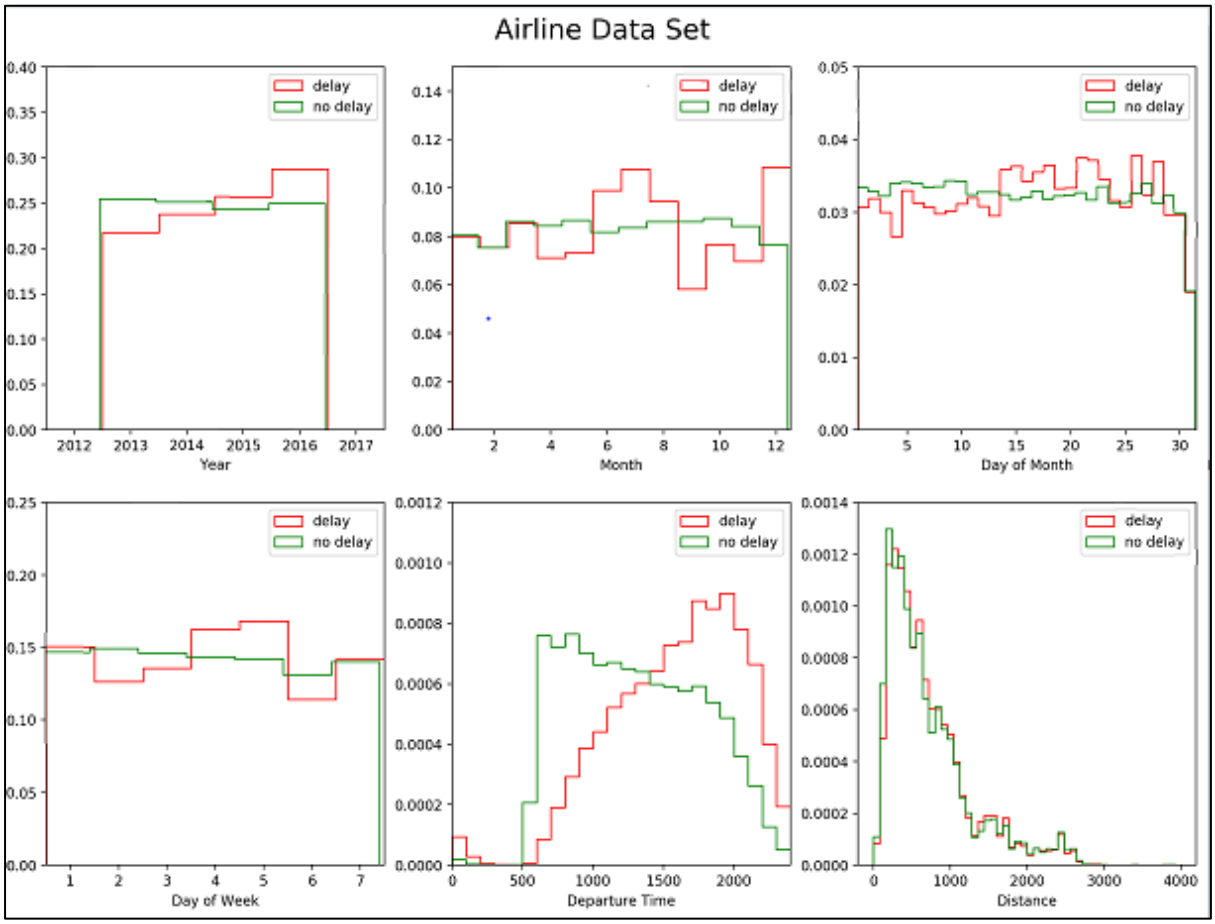
SERIAL NO:	FEATURES	INTRINSIC DISCREPANCY
1	DepTime	0.2400
2	Month	0.0205
3	Distance	0.0114
4	DayOfWeek	0.0069
5	Year	0.0063
6	Origin ORD	0.0058
7	UniqueCarrier	0.0056
8	DayofMonth	0.0050
9	UniqueCarrier	0.0046
10	Origin_HNL	0.0036

From the above table, it can be seen that Departure Time is the most discriminating feature followed by Month and Distance.

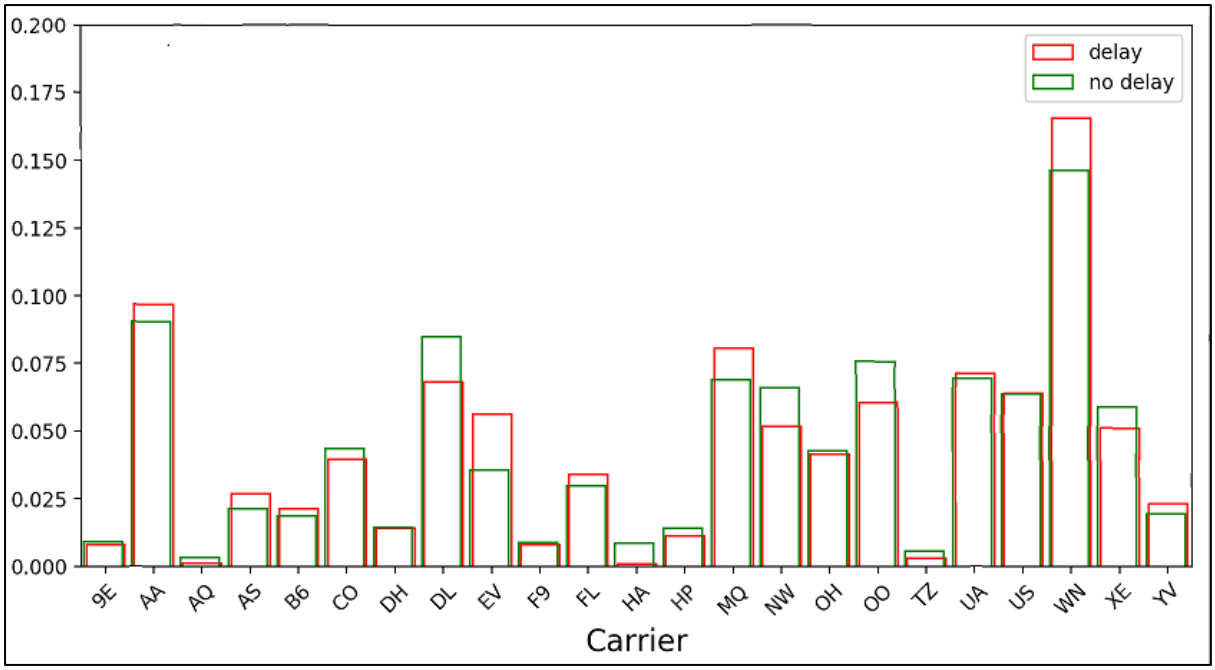
6. EXPLORATORY DATA ANALYSIS

In order to gain insight into the features, we then plot their normalized distributions separately for the “delay” and the “non-delay” cases. Due to normalization, these plots have a statistical relationship where any point along the X axis, the ratio of the two distributions represents a likelihood ratio.

The normalized histogram plots for the different features are plotted as below:



The remaining categorical features like Name of Carrier and Source and Destination airports are represented using bar graphs.



We can observe several important features about the data from the above plots.

- The first important observation that we can make is that delays tend to be more likely during summer months and in December and during the second half of the month and in the middle of the week. We can also infer that delays are more frequent for flights which leave late in the day and that shorter flights are usually less delayed than longer flights.
- The sharpest difference in delay and non-delay is exhibited by the *departure_time* feature.
- We can also observe that flights of some carriers are generally delayed more and similarly flights in and out of some specific airports have more chances of getting delayed.

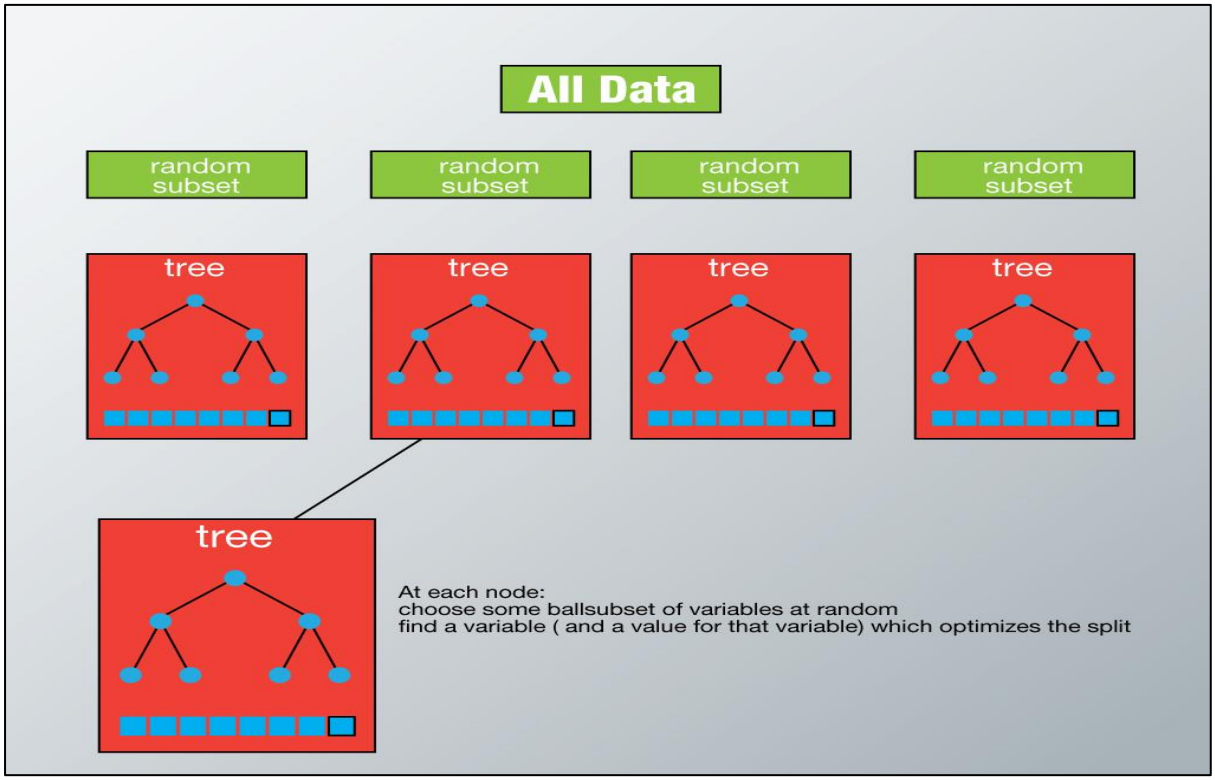
7.MACHINE LEARNING MODEL

For our project we used the Random Forest classifier. Random Forest is chosen as the Machine Learning model because the versatility of the method. Apart from its versatility, this method is capable of doing both classification and regression tasks and it also undertakes dimensionality reduction methods. Hence, we did not do a Principal Component Analysis on the data explicitly.

7.1 RANDOM FORESTS

Random forests or random decision forests[5] are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the

classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.



Our choice of Random Forest is mainly due to the various advantages it offers as compared to other Machine Learning algorithms.

Advantages:

- This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts.

- One of benefits of Random forest which excites me most is, the power of handle large data set with higher dimensionality.

- Random Forest has a property of outputting the importance of features by plotting their relative importance.

7.2. MODEL FITTING:

In this step, we explain the steps we followed to divide the dataset into Training and Testing datasets and how we choose the optimal set of hyperparameters for the Machine Learning model. The steps can be explained as below:

STEPS:

- In the first step, we randomly divided the entire data into Training and Testing data in the ratio of **70% to 30%** respectively.
- Then, we specify a reasonable search grid in the hyperparameter space.
- Then we, perform a 3 – fold Cross Validation on the 70% Training data at each point in the hyperparameter space, to find the optimal set of hyperparameters in the hyperparameter space.
- We used sklearn's GridSearchCV method to search over the hyperparameter space.
- The hyperparameters that we consider here for the Random Forest model are:
  - Number of trees
  - Maximum depth of trees
  - Minimum number of training examples in a newly created leaf

    .
- After performing the Grid Search over the hyperparameter space and performing the 3- fold Cross validation, we obtained an optimal set of hyperparameters as follows:

  - Number of trees: **200**
  - Maximum depth of trees: **20**
  - Minimum number of training examples in a newly created leaf: **2**
- Then, we used the optimal set of hyperparameters obtained in the above step to retrain the Random Forest model again.
- Finally, we test our trained model on the 30% Test data.

OBSERVATIONS:

We observe, that by performing the 3-fold cross validation to find the optimal set of hyperparameters, the overfitting tendency of the model is reduced, and the model generalizes well on new unseen Test data.

8. RESULTS

In this section, we discuss the results that we obtained after our trained model is tested on the Test data. The results are summarized in the below steps.

- Since, in our case, the dataset is heavily imbalanced (**81% data is non-delay**) and (**19% data is delay**), we cannot use absolute percentages to gauge the performance of the model.

- So, we summarise our results using a confusion matrix. A confusion matrix is tabular representation of the results obtained to gauge the performance of a classifier. The prediction results of our trained algorithm on the 30% Test data (30,000 records) can be explained in the below confusion matrix.

	PREDICTED NO DELAY	PREDICTED DELAY	TOTAL
Actual No-Delay	tn = 15458	fp = 8747	24205
Actual Delay	fn = 1881	tp = 3914	5795
Total	17339	12661	30000

From the above confusion matrix, we calculated the various performance metrics for our model as below:

SENSITIVITY (Probability to identify a True Positive): **67.5%**

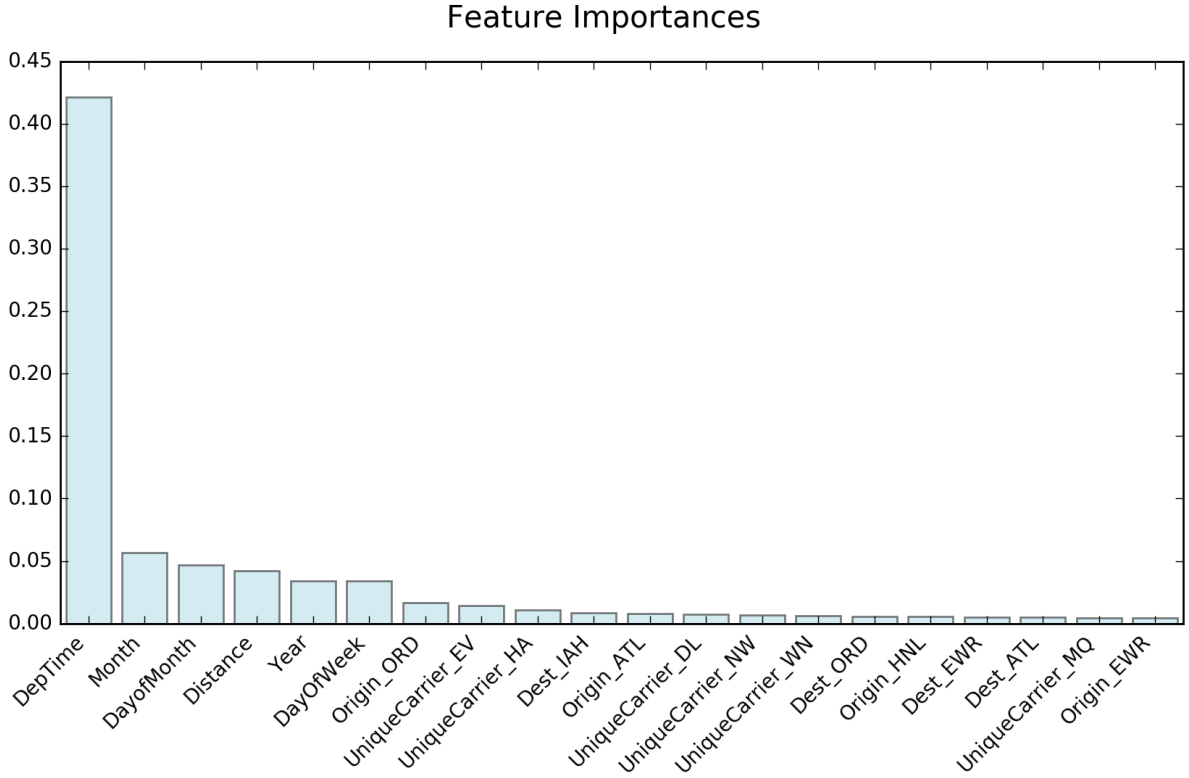
SPECIFICITY (Probability to identify a True Negative): **63.9%**

PRECISION (Probability that a negative id is True): **89.2%**

ACCURACY (Probability of a correct identification): **64.6%**

Note: The above results are calculated relative to a 50% threshold on the output of the classifier.

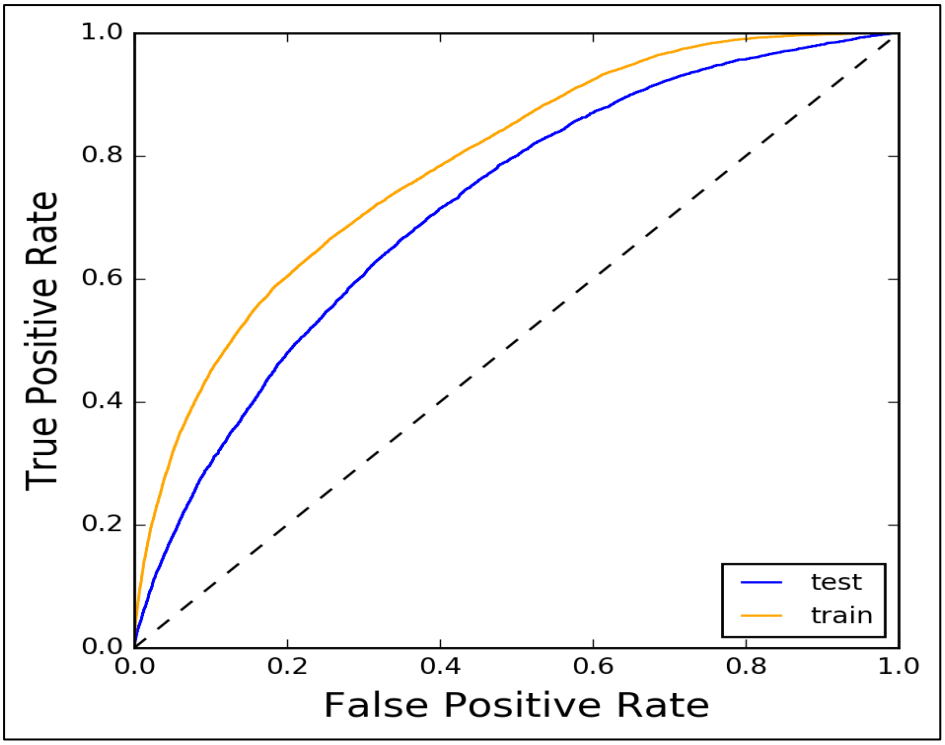
- Then, we use Scikit-learn's inbuilt method by which it plots the most important features for a Random Forest classifier to plot the most important features for our model as shown below:



From, the above plot we, observe that according to Scikit-learn, the most important features for the Random Forest classifier are Departure Time, Month and DayOfMonth which matches with the results we obtained by calculating the intrinsic discrepancies of features in section 8 above.

- Finally, we calculate the area under the Receiver Operating Characteristic(**ROC**) curves for the train data and test data.
  A ROC curve is a curve which measures the diagnostic ability of a classifier as the threshold limit is varied.

  In the below ROC, plot we plot the True Positive Rate (**TPR**) along the Y axis and the False Positive Rate (**FPR**) along the X axis and then measure the Area Under Curve (**AUC**) for the Training dataset and the Testing dataset as the threshold limit is varied.



From the above ROC curve, we calculate the Area under Curve for the training dataset is **0.786** and the Area under Curve for the testing dataset is **0.717.**

Note: We can also see from the above plot, is that the TPR vs FPR curve for the training dataset and the testing dataset are quite close to each other. This gives us a visual confirmation that there is minimum overfitting for our model and justifies the hyperparameter optimization process we performed in the preceding section.

9. CONCLUSION:

In this project we downloaded the publicly available Flight dataset from the Bureau of Transportation Statistics(BTS) website for the year 2013 to 2016 and trained a Random Forest model to predict whether a particular flight will be delayed or not. A flight is considered as delayed if it is arriving more than 15 minutes after its scheduled arrival time, this is the criteria used by FAA (Federal Aviation Administration). Nine features from the dataset was considered to predict the delay. We got an accuracy of 64.60%. We are confident that the result can be improved by incorporating the following points.

- The accuracy of the model can be improved by using more amount of data while training the model. We were able to consider data only from 2013 to 2016 due to the hardware constraints we had.
- Using more data will increase the training and hyper-parameter tuning time substantially. This time can be reduced by using GPUs.
- Also, we can leverage Big Data frameworks like Apache Spark and Hadoop to distribute the training process between cluster of computers. This speeds up the process considerably.

- Accuracy can also be improved by increasing the number of features considered for training. This can be done by appending the weather database to the flights database, thereby increasing the number of features.

REFERENCES:

1 J. J. Rebollo and H. Balakrishnan. Characterization and prediction of air traffic delays. Transportation Research Part C: Emerging Technologies, 44(0):231-241, 2014.

2 Zonglei, Lu, Wang Jiandong, and Zheng Guansheng. "A new method to alarm large scale of flights delay based on machine learning." In Knowledge Acquisition and Modeling, 2008. KAM'08. International Symposium on, pp. 589-592. IEEE, 2008.

3 Zonglei, Lu, Wang Jiandong, and Xu Tao. "A new method for flight delays forecast based on the recommendation system." In Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on, vol. 1, pp. 46-49. IEEE, 2009.

4 Khanmohammadi, Sina, et al. "A systems approach for scheduling aircraft landings in JFK airport." Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on. IEEE, 2014.

5 Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.