PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

A Seminar Report On

Weather induced airline delays prediction using machine learning (random forests)

SUBMITTED BY

Aniruddha Humane

3321

TE 3

GUIDED BY

PROF. A. D. Patil

ISO 9001 : 2008 Certified

COMPUTER ENGINEERING DEPARTMENT

Academic Year:2016-17

PUNE INSTITUTE OF COMPUTER TECHNOLOGY,

DHANKAWADI PUNE-43.

CERTIFICATE

This is to certify that Mr. Aniruddha Sanjay Humane,

Roll No. 3321 a student of T.E. (Computer Engineering Department) Batch 2016-2017, has satisfactorily completed a seminar report on "Weather Induced Airline Delays Prediction Using Machine Learning (Random Forests)" under the guidance of prof. A. D. Patil towards the partial fulfillment of the third-year Computer Engineering Semester II of Pune University.

_____	_____
Prof. A. D. Patil	Dr. R. B. Ingle
Internal Guide	**Head of Department,**
	Computer Engineering

Date :- 30/03/2017

Place :- P.I.C.T., Pune

WEATHER INDUCED AIRLINE DELAYS PREDICTION USING MACHINE LEARNING (RANDOM FORESTS)

Abstract:

The primary goal of model proposed in this seminar is to predict airline delays caused by inclement of weather conditions using data mining and supervised machine learning algorithm (Random Forest). 2008 US domestic flight data and weather data was extracted for training and prediction. Four different models were developed to for analysis of behavior of different parameters. Departure and Arrival delays were separately determined. OOB score was calculated to determine optimum number of trees. Sampling techniques (SMOTE) were then applied on data to improve the performance of model. Every model's performance was compared using precision, recall, F1 score, Accuracy, Confusion matrix, and AUC under ROC.

Keywords*: Data Science, Data mining, Machine Learning, Delay Prediction, Weather, Imbalanced Training data, Sampling Techniques, Binary classification*

INTRODUCTION

With the reduction of operating costs in the aviation industry due to fuel efficient aircraft, the cost reduction by leveraging modern technology and the increase in household disposable income, the volume of air travel increased from 450 billion passenger-miles in 1997 to 600 billion passenger-miles in 2014. (BTS, US Passenger Miles Table) [1]

The higher passenger traffic and the increase in the number of flights offered by airlines, means that during bad weather conditions the National Airspace System (NAS) capacity in the United States is challenged to handle the number of scheduled flights. Passengers can book flights up to one year before the departure date, which is usually when an airline publishes its flight schedule. However, the planned and published flight schedule does not account for the potential impact of weather that may occur on the day of the flight. Instead, the schedule is mainly set according to profit and market share considerations. As a result, the imbalance between flight demand and NAS capacity in the US yields flight delays. The average load factor in a flight for domestic operations in 2012 was close to 83%.

Flight delays not only cause time loss for passengers but also create multiplicative inefficiency, wreaking havoc downstream by disrupting airport runway operations and the planning of airlines. the annual cost of domestic flight delays to the US economy was estimated to be $31-40 billion in 2007 (Joint Economic Committee, US Senate 2008). Correctly predicting flight delays allows passengers to be prepared for the disruption of their journey and allows airlines to pro-actively respond to the potential causes of the flight delay to mitigate their impact.

The abundant research efforts from data scientists, researchers, companies and government agencies on airline flight delays confirms that this is an important area. In particular, the main benefits of better flight prediction are significant operational cost savings and a non-negligible improvement in quality of life for those who use air as an important mean of transport. An accurate online flight delay predictor would certainly generate a lot of interest in the world of air travel.

The goal of work done in this seminar is to use exploratory analysis and to develop machine learning models to predict airline's departure and arrival delays. Based on the literature reviews, this type of problem is actively examined by many researchers and GE even brought out a flight quest challenge with an award of $250,000 to the team who can most accurately predict flight delays.

SURVEY OF MATHEMATICAL MODELS

Sun Choi, Young Jin Kim, Simon Briceno and Dimitri Mavris[2] suggest that, Random forest is an ensemble of many individual decision trees. It builds a large collection of de-correlated trees which are noisy but unbiased, and averages them to reduce the variance. Random forest obtains a class vote from each tree, and then classifies a sample using majority vote. Let $C_b(x)$ be the class prediction of the b^{th} tree, then the class obtained from random forest, $C_{rf}(x)$, is

$$C_{rf}(x) = majority\ vote\ \{C_b(x)\}_1^B$$

However Leo Breiman[3] defines random forest as, A random forest is a classifier consisting of a collection of tree structured classifiers *{h(x, Θk), k=1, ...}* where the *{Θk}* are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x*.

Given an ensemble of classifiers $h_1(x), h_2(x), \ldots, h_K(x)$, and with the training set drawn at random from the distribution of the random vector Y,X, define the margin function as

$$mg(X,Y) = av_k I(hk\ (X)=Y) - max_{j \neq Y}\ av_k I(h_k(X)= j\ ) .$$

In random forests, $h_k(X) = h(X,\Theta_k)$. For a large number of trees, it follows from the Strong Law of Large Numbers and the tree structure that:

As the number of trees increases, for almost surely all sequences $\Theta_1 , \ldots$ *PE** converges to

$$P_{X,Y}(P_\Theta(h(X,\Theta)=Y) - max_{j \neq Y} P_\Theta(h(X,\Theta)= j) < 0)$$

and the strength of the set of classifiers *{h(x,Θ)}* is

$$s=E_{X,Y}mr(X,Y)$$

An upper bound for the generalization error is given by

$$PE^* \leq \rho(1-s^2\ )/\ s^2$$

N. V. Chawla[4] state that, SMOTE is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors.

Algorithm SMOTE (T, N, k)
Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbors k
Output: (N/100)* T synthetic minority class samples
(∗ If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd. ∗)

PROPOSED MATHEMATICAL MODEL

Let S be the system of solution to the problem to predict flight delays, such that,
$S = \{s, e, X, Y, F_{me}, DD, NDD, Su, Fl, \phi_s\}$
Such that,

s = start state = $\{\phi\}$
e = end state = {Y}
X = input set
Y = output set
F_{me} = set of functions
DD = Deterministic Data
NDD = Non-deterministic data
ϕ_s = Constraints

X = {F, W} where
F : Flight data
W : Weather data
Y = {0, 1} such that
0: Flight is on time
1: Flight is delayed

F = {F1, F2, F3}
F1: Function to merge flight and weather data: X1→Y1
X1 ={F_features, W_features}
Y1 = {dataset}
dataset = $\{F_i + W_i \mid F_i \in F[F\_features] \mid W_i \in W[W\_features] \}$

F2: Function to generate data: X2→ Y2
X2 = {dataset}
Y2 = {TrainX, TestX, TrainY, TestY}
d_thresh = 15
dir = "Origin" | "Destination"
if dir is "Origin" then
Xcols = {Features related to origin}
Else
Xcols = {Features related to destination}
X_values = dataset[Xcols]
Y_values = dataset[Ycols]
sm = SMOTE(random_state=42)
X_sampled, Y_sampled = sm.fit_sample(X_values,Y_values)
Split data into TrainX, TestX, TrainY, TestY such that TrainX $\subset$ X_values, TrainY $\subset$ Y_values, TestX $\subset$ X_values, TestY $\subset$ Y_values

F3: Function to create and train classifier: X3→Y3

X3 = { TrainX, TrainY}

Y3 = {pred}

NTrees = 70

clf_rf = RandomForestClassifier(Ntrees)

clf_rf.fit(TrainX, TrainY)

pred = clf_rf.predict(TestX)

F4: Function to calculate confusion matrix: X4→Y4

X4 = {pred, TestY}

Y4 = {TP, TN, FP, FN}

TP = $\{pred_i = TestY_i \mid pred_i , TestY_i = 1\}$

TN = $\{pred_i = TestY_i \mid pred_i , TestY_i = 0\}$

FP = $\{pred_i \neq TestY_i \mid pred_i = 1, TestY_i = 0\}$

FN = $\{pred_i \neq TestY_i \mid pred_i = 0, TestY_i = 1\}$

F5 = Function to calculate accuracy, Precision, Recall, Fscore : X5→Y5

X5 = {TP, TN, FP, FN, pred, TestY }

Y5 = {Accuracy, Precision, Recall, Fscore}

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

Fscore = 2*Precision*Recall/(Precision+Recall)

Accuracy = (pred == TestY)/count(pred)

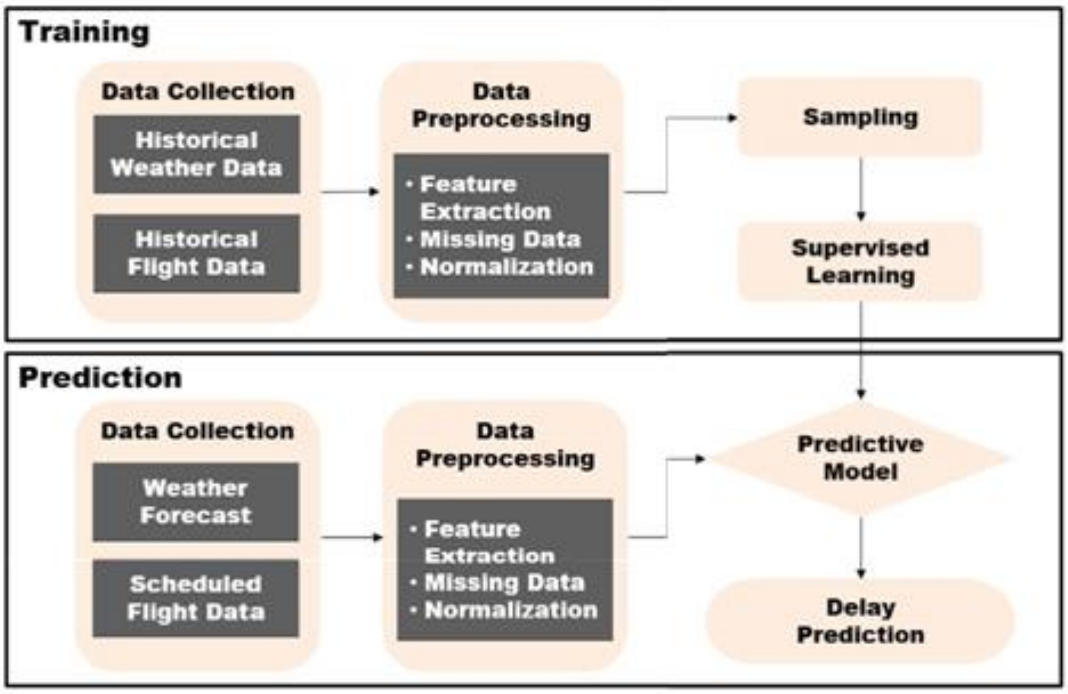
DD = {F,W}

NDD = {NaN, Invalid data}

DESIGN AND ANALYSIS OF SYSTEM

a. Analysis Process Flow

b. Prediction Model



c. Data Acquisition
   1. The flight data, also known as on-time performance data can be downloaded from the American Statistical Association.
   2. Historical weather data and flight demand data for 2008 is from the FAA Aviation Systems Performance Metrics (ASPM).

d. Libraries used:
   1. Python/ipython
   2. Numpy
   3. Matplotlib
   4. Scikit-Learn
   5. Scipy
   6. Pandas
   7. Imbalanced-Learn

e. Environment:

1.	Processor	:	Intel i5-5200U 2.20GHz Dual core
2.	RAM	:	8GB
3.	Operating System	:	Windows 10 Pro
4.	Language platform	:	iPython

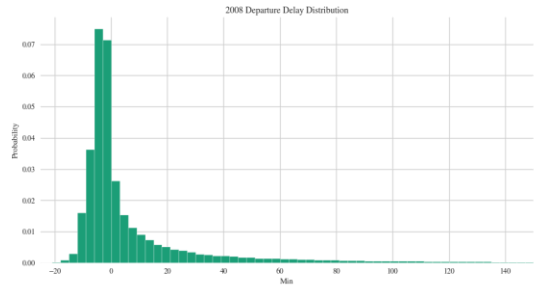
f. Data Preprocessing
   1. US 2008 flight and weather data had almost 84 features with about 7 million data samples.
   2. Hence, only features which are important for Delay Prediction were extracted and rest of the features were ignored.
   3. Some features had invalid values like NaN these samples were removed
   4. As it is easier to process numbers than strings, features with string values were factorized and converted to numerical values

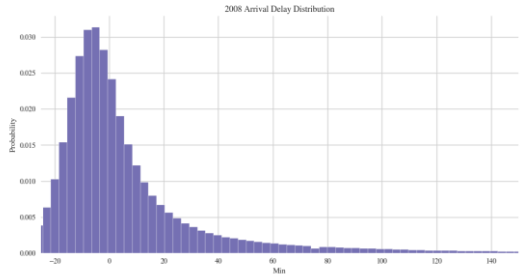
g. Prediction Models
   1. Analysis on 4 different models
   2. Prediction performance was compared with the help of Accuracy, Precision, Recall, F1 score, Confusion Matrix and AUC under ROC

DISCUSSION ON IMPLEMENTATION RESULTS

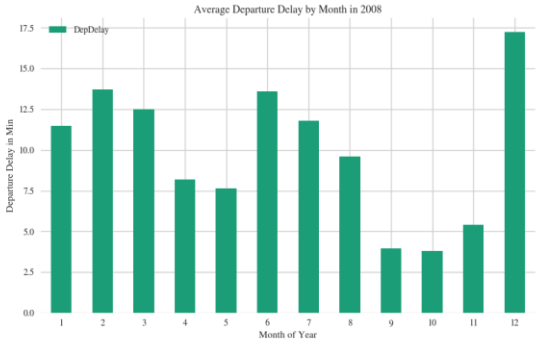
Data Analysis

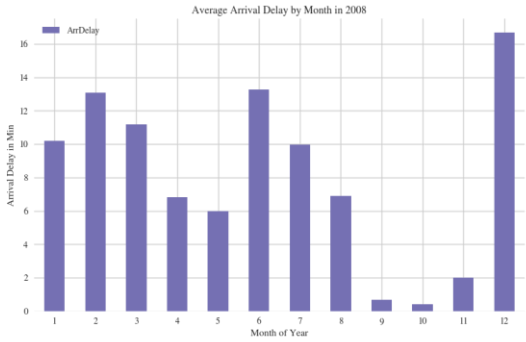
a. Departure and Arrival Delay Distribution with Respect To Amount Of Delay



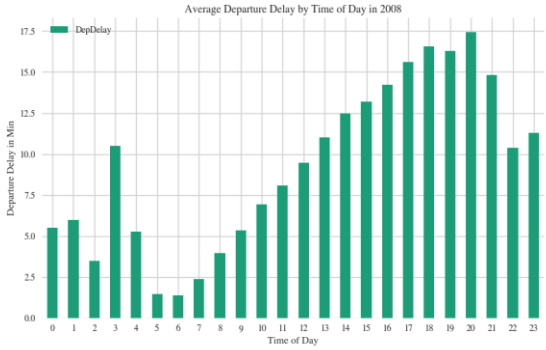


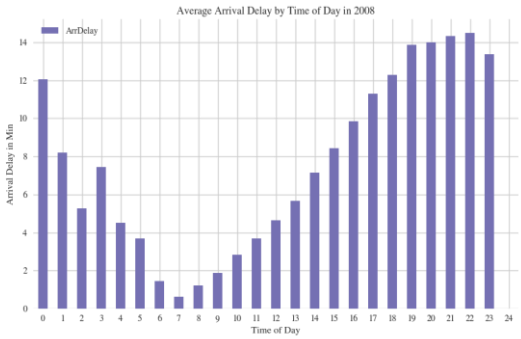
b. Departure and Arrival Delay Distribution With respect to Months



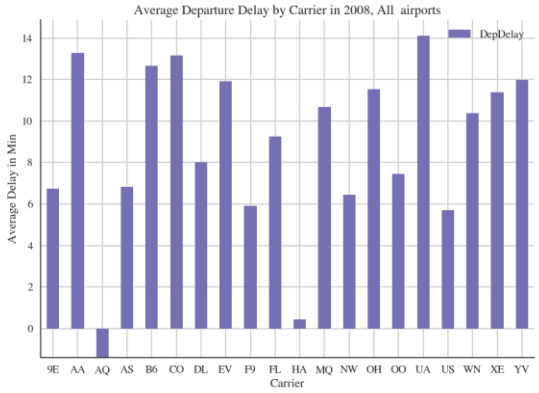


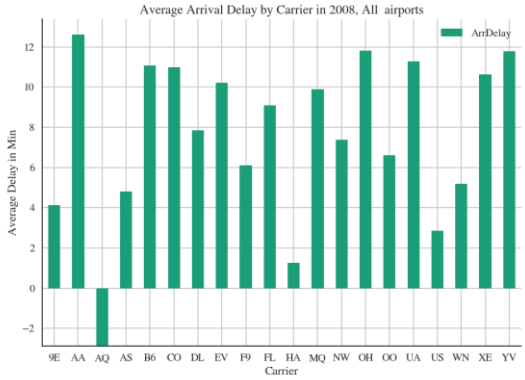
c. Departure and Arrival Delay Distribution With respect to Time of day





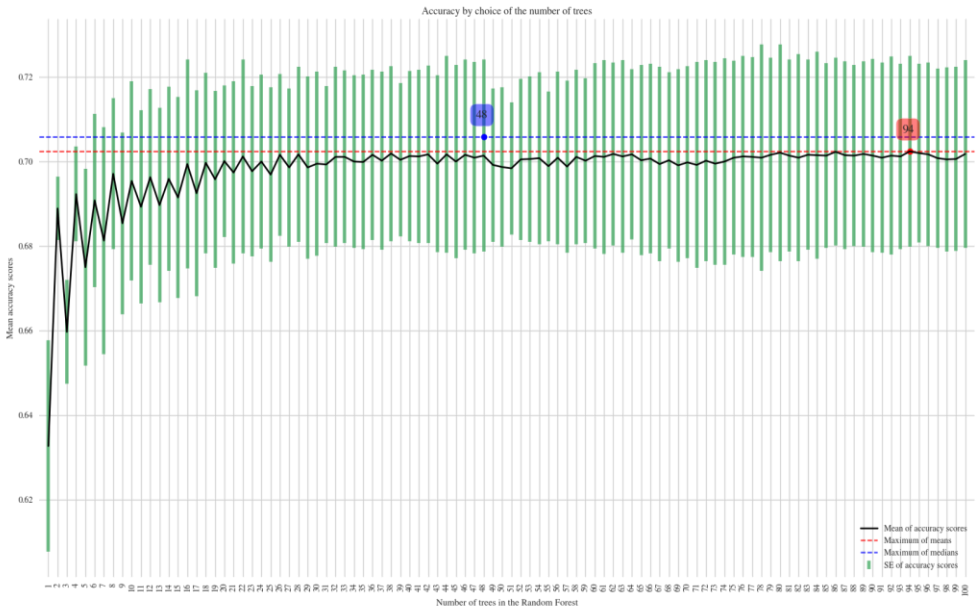
d. Departure and Arrival Delay Distribution With respect to Carrier



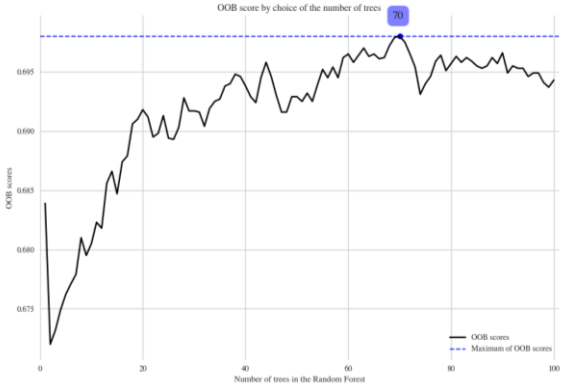


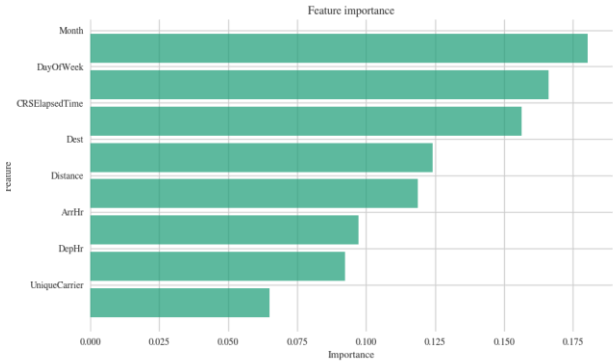
Prediction models: Four models were made

1. Model to predict flight departure delays from Chicago O'Hare International Airport (ORD)
    a. Random 20k samples were chosen
    b. Maximum features selected = 2
    c. Classifier was iterated with no of trees from 1 to 100 and performance of every tree wrt mean accuracy was plotted ($mean_{max}$ = 94, $median_{max}$ = 48)

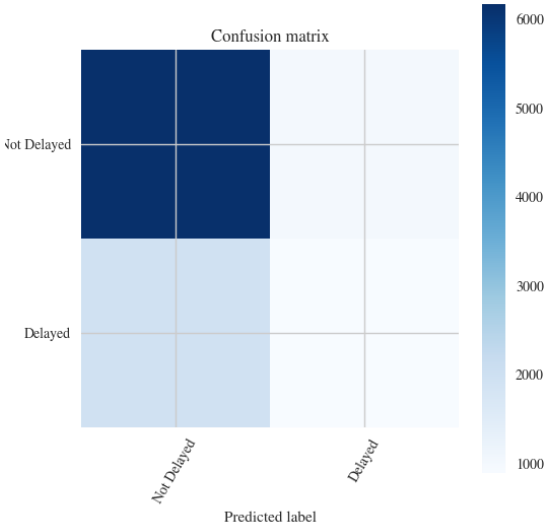


d. OOB score for this model was as follows (NTrees = 70) and feature importances





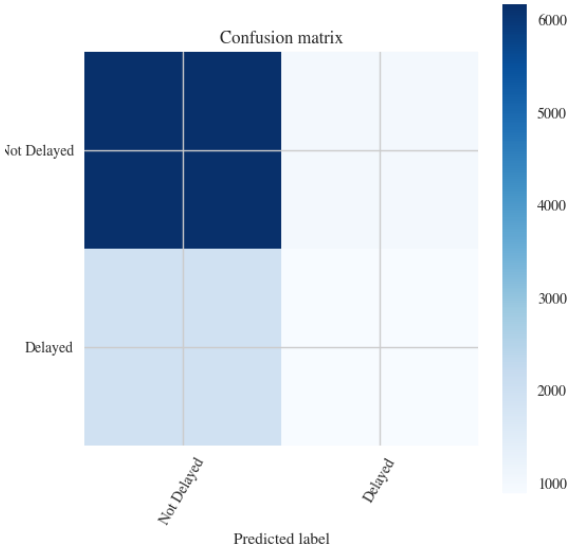
e. Confusion matrix and overall performance attributes (Accuracy = 71%)



Precision	47%
Recall	32%
F1 score	38%
Accuracy	71%

	On time	Delayed
On time	6170	1000
Delayed	1934	896

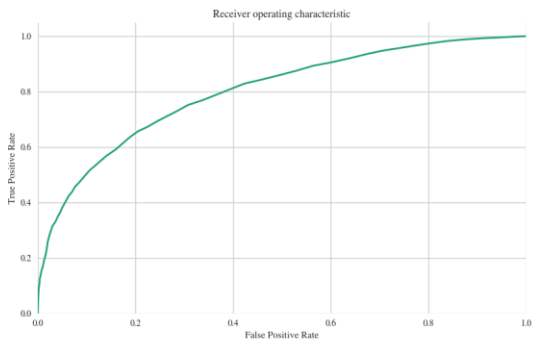
2. Model to predict departure delays using subset of weather and flight data from Chicago O'Hare International Airport (ORD)
   a. Performance and confusion matrix calculation with NTrees = 70

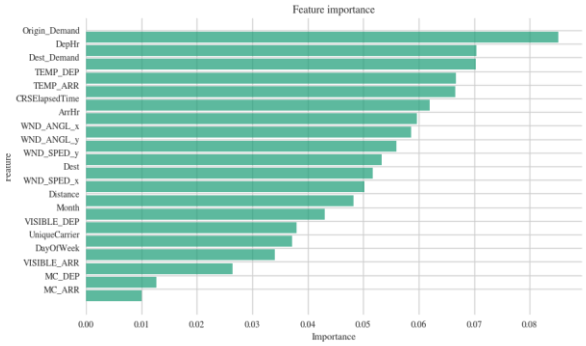


Precision	72%
Recall	50%
F1 score	59%
Accuracy	76%

	On time	Delayed
On time	5929	663
Delayed	1706	1702

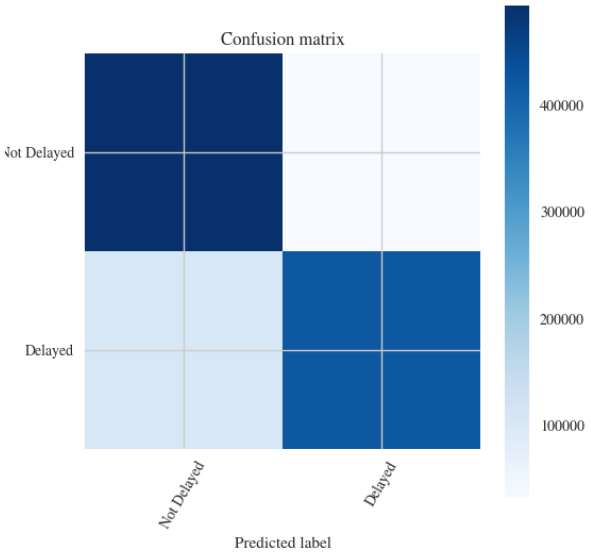
b. ROC of model (Area Under Curve = 0.80) and New Feature Importances





3. Model to predict Departure delays using whole weather and flight data from Chicago O'Hare International Airport (ORD)
    a. With NTrees = 70 the classifier model was trained and tested with whole dataset
    b. Confusion matrix in TP region was much darker indicating that this model was predicting more delayed samples than previous model.
    c. However in this case adding more samples did not help.
    d. This model was significantly affected with accuracy paradox.
    e. Performance did not improve that much with accuracy of 75%
    f. Both accuracy and precision were decreased by 2%.
    g. Even AUC was decrased to 0.76.

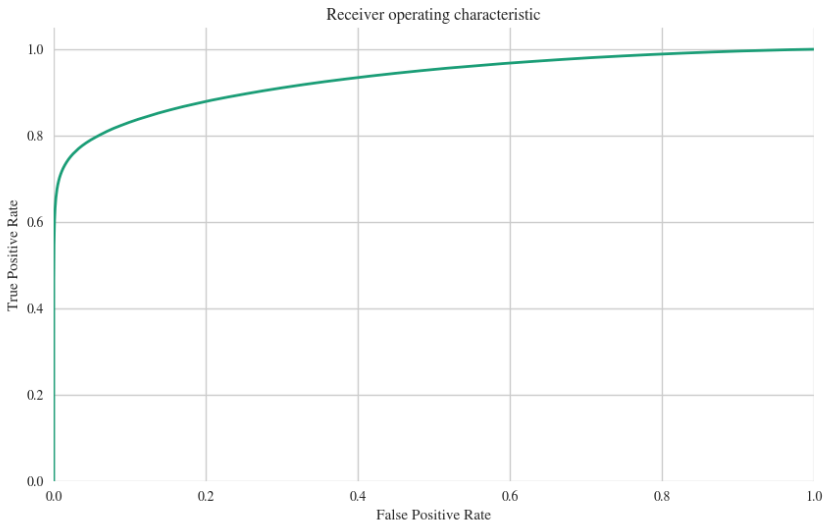
4. Model to predict Departure delays from Chicago O'Hare International Airport (ORD) using weather and flight data with sampling.
    a. Model was trained with same parameters, However here, sampled data was used
    b. Confusion matrix and performance calculation results are as follows



Precision	93%
Recall	80%
F1 score	86%
Accuracy	87%

	On time	Delayed
On time	493780	32630
Delayed	103239	422192

c. Receiver Operating Characterics and Area Under Curve was 0.93



CONCLUSION AND FUTURE ENHANCEMENT

From the analysis done on Flight data, The delay distribution were centered around zero. Some flights were delayed by more then 2 hours. December has largest amount of delays mainly due to snowstorms also June and July have delays due to summer vaccations. October and november are the months with least amount of delays. we see a marked "V" shaped decline in delay with the lowest delays in early morning hours. Both departure and arrival delays accumulate from the earlier morning hours reaching their peaks in the evening hours indicating Flight Delay Propagation.

On execution of first model, with random 20k samples, By trial and error and Accuracy and OOB score based on no of trees, Optimum number of trees was found out to be 70. Month played a vital role in flight delays along with Day, Duration and Destination. Maximum accuracy of 72% was obtained which was encouraging. However Precision and F1 score needed improvements

Hence, weather data was integrated with flight data. Final dataset had 34 features with 1.5 million records. 25% improvement in precision along with 21% improvement in F1 score was observed. Maximum accuracy given by this model was 79% with 18% increase in delayed flights prediction. But on training this model with all available data samples did not improve performance of prediction. This model was predicting Non delayed flights 88% of the time hence showing accuracy paradox.

To overcome this, Sampling techniques (SMOTE) were applied to whole dataset. With increased number of minority samples. The Random Forest Classifier was Trained and tested again on this data. This model was performing outstandingly well. This model was predicting delayed flights 88% of the time. A significant improvement in Precision of 93% and F1 score of 86% was obtained. AUC of ROC was 0.93 Final accuracy of model was found out to be 87%. Hence improvement in performance of model by oversampling minority delay samples was non negligible.

Performnce of this model can be improved by using mode number of samples, by using more number of trees. However more number of trees means more execution time. Hence optimum number of trees should be used. More features which can give detailed information about flight delays can be used to improve performance even further.

REFERENCES

[1] BTS, US Passenger Miles Table

[2] Sun Choi, Young Jin Kim, Simon Briceno and Dimitri Mavris - Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms Georgia 30332–0250

[3] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001

[4] N. V. Chawla, "C4.5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure," in Proceedings of the ICML, Workshop on Learning from Imbalanced Datasets II, Washington DC, 2003.

[5] S. S. P. Reshma C. Bhagat, "Enhanced smote algorithm for classification of imbalanced big-data using random forest," in Proceedings of the Advance Computing Conference (IACC), 2015.