International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Weather-based Aviation Delay Predictions

Dr. K Sreekumar, Varun Iyer, Karan Matalia

Abstract*: *Airline Industry is one of the major contributors towards socio-economic utilityand forms a vital part of worldwide transportation system. The aviation industry hasevolved immensely over the past couple of decades, however flight delays and cancellations are an inevitable part of the industry thathurts passengers along with the airlines and the airport itself. Given flight delays results in economic and environmental impact, therefore, it becomes absolutely essential to improve the air traffic management. In thispaperwepredictflightdelaysincluding delay due to factors likeinclement weather conditions,precipitation, temperature. Wind speedetc. We prediction models by leveraging the power of data science and machine learning models.It was carried out in two parts – the first being classification and the second using multiple regression techniques and evaluated the best models based on their respective accuracy parameters. Data Scientist have lately found excellent results using Gradient Boosting and hence we intend to utilize it for better accuracy and predictions. Our work could be used by the airliners and the passengers alike for getting an idea about the delays so that they can effectively manage their time and resources beforehand hence minimizing wastage.

***Index Terms:*Decision Tree, Gradient Boost,Passenger Carrier on-time performance,Weather, XGBoost**

I. INTRODUCTION

Today majority of the world's population are increasingly moving towards air transportation for theircommute.Flightdelaysareaninexorablepart oftheaviationindustrythatresultsinlossofmulti-billiondollarsforairlines.Inaresearchby'Bureau of Transportation Statistics (BTS)' , it was discovered right around 20 percent of all the planned business flights confronted delays. Moreover it is roughly calculated that flightdelays account to a loss of a staggering $41 billion every year to the United States nationaleconomy [1].It was reported that approximately a loss of $33 billion was incurred to flyers, aviation companies and related areas of National Airspace System during the year 2007 , both directly and indirectly. Also if we consider the environmental impacts of flightdelaysitwasevaluatedthattaxi-outexercises added 4,000 tons of hydrocarbons, twofold the nitrogen oxides and an enormous 45,000 tons of carbon monoxide emanations in the United States in 2007[2].

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Varun Iyer*, Department of CSE, SRM Institute of Science & Technology, Chennai, India
Karan Matalia, Department of CSE, SRM Institute of Science & Technology, Chennai, India.
Dr.K. Sreekumar, Department of CSE, SRM Institute of Science & Technology, Chennai, India.

Since customer base, loyalty, frequent-flyer rewards etc. are also affected due to the drop in the performance standards due to delays, therefore, delays also endangers airliner's reputation. Therefore prediction of airline delays can help the airliners as well as passengers to effectively inform passengers as well as the airlinerswellinadvancetoproductivelyreorganize their plans optimization.and operations for economic viability and time resourceThere are 5 primary caused that aviation delays have been categorized into by the Bureau of Transportation Statistics namely air carrier, security, late-arriving aircraft, National Aviation System and extreme weather [3]. Apart from weather , other causes also constitute a major chunk towards flight delays , however weather is the prime factor for flight delays and cancellation apart from all other delays. Also it directly or indirectly affects other factors of delay , for instance, inclement weather condition might force Air Traffic Control to re-route flight pathsand trajectory of certain flights for safety and logistic convenience. Keeping in account the statistics and factors, weather accounts for over 40% of total delay time(in minutes) in the aviation industry[3].Hence it becomes essential to predict flight delays using weather as a feature for high accuracy. Hence, through this paper we complement the existingresearchforflightdelay arrival withanemphasis on delay due to weather hence aiding airports and airliners to effectively adjust their economic resources and manpower.

II. RELATED WORK

S. Choi et.al [4] primarily proposes a system to predict flight delays that are predominantly caused by inclement weather conditions . The author uses supervised machine learning algorithms for this purpose.It was concluded that by using supervisedmachine learning algorithms along with effective use of SMOTE flight delay was successfully predicted howeveruncertaintyintheforecastwouldimprove would enhance themodel.Balakrishnan, H. et.al [5] perform a comparative analysis of Machine Learning Models in order to predict delays in the Air Traffic Network. The author shows the variance in the performance of the of the Markov Jump Linear System(MJLS), classical machine learning techniques like three candidate Artificial Neural Network (ANN) architectures and Classification and Regression Trees (CART). KuhnN.et.al[6]usesalgorithmslikedecisiontree and other such Machine Learning Algorithms for flight delay predictions and attains a test accuracy ofover91%forthegivenclassifiers Inthispaper,all those datasets are considered which have over 15 minutes of difference between scheduled and actual arrival time .

The end result was that the test accuracies were approximately 91% wherein the depth of the treewassevenandtotalleafnodeswere127which constitutedless than one percent of the total examples that was trained. MovvaN.et.al[7]researchedonflightdelaysdue to sub-optimal weather conditions . They use datasets from US Bureau of Transport Statistics with Local Clamatorial Data fromNOAA.They created the training, the validation and the testing splits of the data in 60%-30%-10% ratio.They tuned l2-regularization for XGBoost and L1-regularization strength for Lasso. The XGBoost model's recall turned out to be 15.5% with a 70% precision.V.C Kotak et.al [8] in their paper use the OneR algorithm which is a one-level decision treein order to predict the delays . The accuracy for Ibk and OneR gave the same accuracy as that ofNaïve Bayes and Bayes at around 54.37% and 54.81% respectively with the training time for IBk being more. Manna S. et.al [9]in their paper used gradient boosted decision tree in order to carry out their predictions as it deemed to be very effectual to carry out regression job. The algorithm uses learners that are comparatively frail in order to lowerthevarianceandbias.RandomForestsleads to the reduction in variance , hence it was deemed inappropriate. They used data from Department of Transportationandtheendresultwastheyachieved92.3 percent and 94.85 percent as the Coefficient of Determination for arrival and Departure respectively[9].

III. METHODOLOGY

Our primaryobjectivewastostudythearrival patterns across various airports of various airlines. Then we deployed various. However, majority of flights have almost negligible or little delay inarrival time. Hence, we divided our study into two parts : One to analyze the flight delays for delays less than a couple of minutes and the other for flight delays greater than 5 minute. We deployed multiple machine learning algorithms and trained them to find the best performing Machine Learning Technique.We also deployed Oversampling, which help us give an improvement in our predictions in comparison to under sampling that provides not so accurate results.One of the major oversampling techniques is the SMOTE. SMOTE works by creating additional synthetic dataset samples and creates new minority data in the respective cases hence giving us an improvement in the dataset quality.[15]

A. Supervised Learning Algorithms

In this experiment 5 different algorithms were used, and their subsequent performance was evaluated.

1. Logistic Regression

Logistic Regression is a classification algorithm that has the following hypothesis:

$$h_\theta(x) = \ g(\theta^T x) = \frac{1}{1 + \ e^{-\theta^T x}}$$

Where $\theta^T x = \ \theta_0 + \sum_{j=1}^{n} \theta_x \, x_j$. The parameter θ can be deciphered by maximum likelihood estimation.

2. Linear Regression

Linear Regression is a technique that is used to find the linear relationship between the target values. Regression of *y* on *x* is indicative of the prediction of value of *y* given the values of *x*. The equation is given by

$$y_i \ = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_i} + \boldsymbol{\varepsilon_i} (2)$$

Where y_i is the dependent variable ,β_0 population is the population Y intercept , β_1 is the Population Slope Coefficient and $\varepsilon_i$random error component.

3. Random Forest Regressor

Random Forest Regressor is an ensemble technique that utilizes multiple individual decision trees.It creates a collection of unbiased and de-correlated trees but noisy and reduces the variance by averaging them. The mathematical interpretation is :

$$\overline{r_n}(X, D_n) = \boldsymbol{E_\Theta} \, [\, \boldsymbol{r_n}(\, \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{D_n}\,)] \ (3)$$

Here E_Θ is the random parameter expectation conditionally on X and D_nDataset [11]. Θ is the randomizing variable which is used to justify how successive cuts takes place while the individual trees are being built.

4. Decision Tree

A Decision Tree involves a tree structures resembling flow chart wherein each internal node represents input attributes, the branches represent the outcomes and terminal node receives all examples from the training set. Decision tree learning utilizes a decision tree as a prescient model which maps the findings about a feature to the closure of its target value objective.The tree is partitioned on every node based on the response of the true and false question for the features.Indices like Gini Index or entropy are used quantify impurities with specific nodes. The formula below show how entropy is calculated:

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i \qquad (4)$$

Where S is the current group which we are interested in and p_i is the probability of finding the system in the i^{th} state [12].

5. Gradient Boosting

We primarily use the eXtreme Gradient Boosting (XGBoost) Technique which is an implementation of Gradient Boosted Decision Trees. XGBoost are known for their speed of execution and the performance of the model and hence XGBoost is also used by CERN for classifying signals in the Large Hydron Collider [13]. It is based on the Gradient Boosting algorithm wherein the models predict the errors form the previous models which is then combined together to create the final model.

When the new models are combined, the subsequent losses are minimized using the Gradient Descent Approach and hence the name Gradient Boosting. Random Forests executes its operation of error reduction by decrementing variance whereas XGBoost uses weak learners for error reduction hence performing better than the former[9]. The statistical equation is:

$$F(x, \beta, \alpha) = \sum_{i=1}^{n} \beta_i h(x, \alpha_i) \quad (5)$$

Where h are the weak learners, βi are the weights of each weak learners. The Gradient Boosted Decision trees sequentially grows the trees and re-evaluates the weights of each learner toward the final prediction.

B. Model Prediction

The average arrival delays for are evaluated using two parameters: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

1. Mean Absolute Error

Mean Absolute Error is the mean of every prediction error's absolute values for the given instances of the test data set and mathematically it is the difference between two continuous variables. It can be denoted by:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|r_n - \hat{r}_n| \quad (6)$$

Where $\hat{r}$ is the prediction rating; r_n is true rating in testing data set; Nis the number of rating prediction pairs between the testingdata and prediction result [14].

2. Root Mean Square Error

Root Mean Square Error is the square root of the mean of the squares of all the errors of the given dataset. It is the standard deviation of predictive errors. RMSE is the calculation of how stretched the prediction errors are. It is commonly used in the regression analysis as it helps us to liquidate large errors. The equation is [14]:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(r_n - \hat{r}_n)^2} \quad (7)$$

Where the denotation is same as that of MAE.

3. Precision, Recall and F1 Score

Accuracy does not give us the precise information hence we use a combination of Precision, Recall& F1 score in order to get the accurate score . Precision gives us an indication on how well the model is able to predict the positives. Recall gives us the count of all those actual positives which are categorized as positives. The F1-Score is combination of Recall and Precision and hence it gives us a combination of both of these parameters for an unbiased result.

$$F1\ Score = \frac{2 * Precition * Recall}{Precision + Recall} \quad (8)$$

IV. Data Analysis

A. Data Description

For training our data, we utilized the on-time performance dataset that is publicly available on the United States Department of Transportation for the United States domestic air-traffic taken from TransData. The data set is primarily from the month of January 2018 to the month of November 2018. The datasets consists of over 6 million examples in addition to the 27 features namely : Day of Week, Flight Date, Unique Carrier Code, Carrier, Carrier Flight Number, Origin , Destination, CRS Depature Time, Departure Time, Departure Delay, Taxi Out Time, Wheels On Time, Taxi In Time, CRS Arrival Time, Arrival Time, Arrival Delay, Cancelled Status, Cancellation Code, CRS Elapsed Time, Actual Elapsed Time, Air Time, Distance, Carrier Delay, Weather Delay, NAS Delay, Security Delay & Late Aircraft Delay. For weather dataset ordered the Local Climatological Data provided by the of the United States Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) for airports across the United States. The datasets consisted over 90 features out of which we extracted the required features from the given datasets.

B. Data Preprocessing

The tools that were used for this experiment were Pandas, Python, Jupyter Notebook within a condaenvironment. The completeness of both the datasets were checked to eliminate the missing data. The missing data was filled while rows with missing features like flight number etc. which could not be filled due to lack of substantial evidence were removed from the dataset.

Since the dataset was magnanimous therefore the top 10 airports in terms of traffic were extracted from the datasets followed by the visualization for better understanding. The airline datasets were amalgamated with the weather datasets in accordance to the airport locations along with the modification of labels for better optimization and understanding.

TABLE 1. FEATURE DESCRIPTION

No	*Feature Name*	*Feature Description*
1	DAY_OF_WEEK	Indicates the day of week from Sunday to Saturday
2	UNIQUE_CARRIER	IATA assigned code for carrier identification
3	ORIGIN	Origin Airport
4	DEST	Destination Airport
5	ARR_DELAY	The delay in arrival in minutes
6	DEP_HOUR	Hour of Departure
7	ARR_HOUR	Hour of Arrival
8	DEP_HOURLYVISIBILITY	The visibility at Departure Airport
9	DEP_HOURLYDRYBULBTEMPC	The Temperature at Departure Airport
10	DEP_HOURLYWindSpeed	The Wind Speed of Departure Airport
11	DEP_HOURLYPrecip	The precipitation of Departure aiport

12	ARR_HOURLYVISIBILITY	The visibility at Arrival Airport
13	ARR_HOURLYDRYBULBTEMPC	The Temperature at Arrival Airport
14	ARR_HOURLYWindSpeed	The Wind Speed of Arrival Airport
15	ARR_HOURLYPrecip	The precipitation of Arrival airport

C. Analysis of features

The dataset provides a lot of information for to help us understand the flight delay statistics . Fig 1. gives us an idea on the top 10 of the most busiest airport in the United States in terms of traffic . It was found out that airport with IATA codes ATL , ORD & DFW were the most busiest airport in the United States.

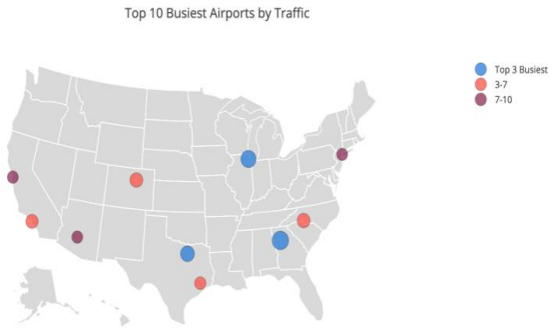

Fig.1 Top 10 Busiest Airport by Traffic

Furthermore we found out that the highest percentage of delayed flights was from JetBlue Airways followed by Frontier Airlines as shown in Fig.2

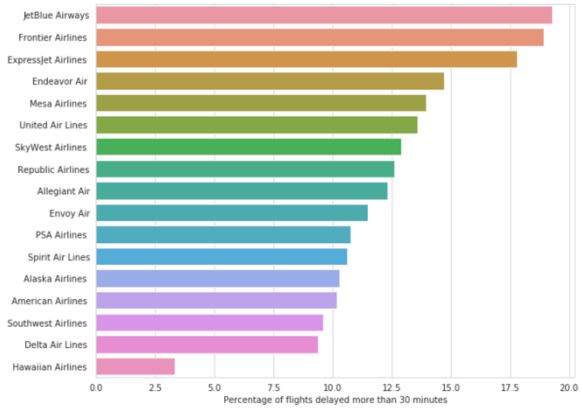

Fig.2 Percentage of Delayed Flights

V. EXPERIMENT

A. Model Construction

The XGBoost based on the Gradient Boosted Decision Tree was selected for our problem. Since we had many numerical features hence dummy variables were created using one hot encoding technique. Nowourprimaryobjectivewastostudythearrival patterns across various airports of various airlines.The following hyperparameters were used for our XGBoost algorithm: 'DAY_OF_WEEK','FL_DATE', 'UNIQUE_CARRIER','ORIGIN', 'DEST', 'ARR_DELAY', 'CANCELLED','DISTANCE', 'DEP_HOUR', 'ARR_HOUR', 'MONTH','DEP_HOURLYVISIBILITY','DEP_HOURLYDRYBULBTEMPC','DEP_HOURLYWindSpeed','DEP_HOURLYPrecip','ARR_HOURLYVISIBILITY','ARR_HOURLYDRYBULBTEMPC','ARR_HOURLYWindSpeed'& 'ARR_HOURLYPrecip'. One-Hot encoding was performed for converting necessary features into categorical variables. However, majority of flights had almost negligible or little delay inarrival time. Hence the analysis for flights with delays less than acouple of minutes was carried out using classification.Subsequently regression model was applied for flights delayed greater than 5 minutes. Linear Regression, Random Forests and Decision Tree Regression algorithm was also deployed for delays greater than 5 minutesin order to cross verify the efficiency of XGBoost. In order to split the dataset into randomly generated training and testing subsets, the *train_test_split*function of from scikit learn was utilized.

B. Comparision and Evaluation

On executing the logistic regression for flight delays less than 5 minutes. We randomly selected over 100,000 samples from the positive and the negative samples in the ratio of 40:60. We averaged our predictions over 100 models and we got an average accuracy of 70.81 % . The cost of false positive and flase negative were also considered in the experiment.

TABLE 2. REGRESSION OUTPUT DESCRIPTION

False Positive	13004
False Negative	16185
True Positive	18426
True Negative	52385

We also found out that the Random Forest Classification Algorithm provided varied results with SMOTE technique and we got results with a precision of 97.88 % as compared to 98.36% of that of without oversampling.The detailed report for classification with SMOTE can be found below.

TABLE 3. RANDOM FOREST CLASSIFICATION OUTPUT

Technique	Precision	Recall	F1 Score
Clasification with SMOTE	0.9836	0.9836	0.98357
Classification without SMOTE	0.9788	0.9789	0.97887

Further we went ahead and performed the various Machine Learning algorithms on flights with delay greater than 5 minutes. We took the logarithmic of the continuous target variable so that any abnormal values in our dataset can be normalized. The following result was obtained for following statistics:

TABLE 4.OUTPUT DESCRIPTION

Algorithm	Mean Absolute Error	Root Mean Square Error
Linear Regression	27.865	66.788

Random Forest	29.894	70.263
Decision Tree	29.985	70.432
XGBoost	6.59	13.32

Here the maximum depth of Random Forest was 5. And had 3000 trees. Decision Tree a maximum depth of 5. XGBoost had 3000 trees with a maximum depth of 5 and a learning rate of 0.1.

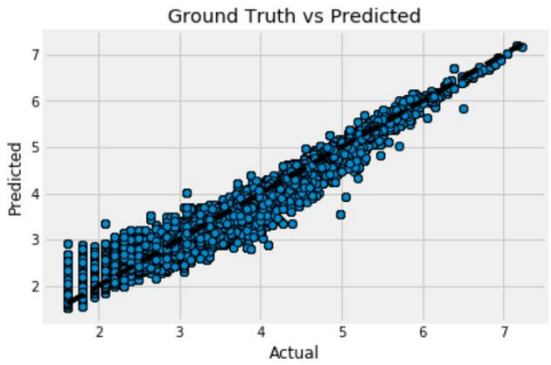

Fig 3. Cross Validated Predictions

VI. CONCLUSION AND FUTURE SCOPE

We evaluated the results in two stages.The data was preprocessed, the flight and weather data were merged , analyzed and then the model was trained for classification and regression. It was found out that Hartsfield–Jackson Atlanta International Airport was the busiest airport and JetBlue had the worst on-time performance during our limited dataset. We got an accuracy of around 70.81% and we got an accuracy of around 97.8% after necessary oversamplingfor classification RMSE of 13 minutes along with an MAE of 6 minutes using XGBoost was obtained for regression which were respectively the best analysis algorithms. Hence it was found out that there are various machine learning paradigms for flight delay predictions. However Gradient Boosting technique performed better than all the other models for regression . This model can be used by passengers for planning their itinerary beforehand and commercial airliners could use it to predict delays and regulate their revenues for maximizing profit, however computational requirements remains a challenge. The experiment was conducted for a period of 9 months due to computational limitations with just the top 10 airports. With data available for the past 20 years for over 500 airports and analyzing all the data would give us accurate results. MJLSalgorithm looks very promising however further research in that field will tell us accurately about its capability. With extended time and computing power we would be interested to set up clusters and run larger models.More generally , we have shown that Gradient Boost for Regression and Random Forests for classification are the best possible method to approach this problem and that the local weather data is indeed a source for predicting flight delays.

REFERENCES

1. H. Balakrishnan, “Control and optimization algorithms for air transportation systems,” Annual Reviews in Control, vol. 41, pp. 39–46, 2016.
2. R. R. Clew low, I. Simaiakis, and H. Balakrishnan, “Impact of arrivals on departure taxi operations at airports,”2010.
3. Airline On-Time Performance and Causes of Flight Delays. (2018, October 09). Retrieved January 15, 2019,from https://www.bts.gov/topics/airlines-and-airports/airline-time performance-and-causes-flight- delays
4. S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, 2016, pp. 1-6. doi: 10.1109/DASC.2016.7777956
5. Gopalakrishnan, K., & Balakrishnan, H. (2017). A Comparative Analysis of Models for Predicting Delays in Air TrafficNetworks.
6. Kuhn, N., & Jamadagni, N. (2017, Autumn). Application of Machine Learning Algorithms to Predict Flight Arrival Delays. Retrieved fromhttp://cs229.stanford.edu/proj2017/final- reports/5243248.pdf
7. Movva,N.,&Menon,S.(2016).Predictingflight delays and cancellations using weather as a feature. Retrieved from http://cs229.stanford.edu/proj2016/report/MenonMovva-PredictingFlightDelays-report.pdf
8. V.C. Kotak5S. O. S. S. (2017). Flight Delay Prediction System Using Weighted Multiple Linear Regression. International Journal of Engineering and Computer Science, 4(04). Retreived fromhttp://www.ijecs.in/index.php/ijecs/article/view/1764
9. S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta and S. Barman, "A statistical approach to predict flightdelay using gradient boosted decision tree," *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*, Chennai, 2017, pp.1-5.doi: 10.1109/ICCIDS.2017.8272656
10. S. (n.d.). Regression Analysis. Retrieved January 24, 2019, from http://home.iitk.ac.in/~shalab/regression/Chapter2-Regression-SimpleLinearRegressionAnalysis.pdf
11. Biau, G. (2012). *Analysis of a Random Forests Model,13*. Retrieved February 04, 2019, from http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf
12. Gupta, Bhupendra and Pauri Garhwal Uttarakhand. “Analysis of Various Decision Tree Algorithms for Classification in Data Mining.” (2017).
13. Chen, T. & He, T.. (2015). Higgs Boson Discovery with Boosted Trees. Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning, in PMLR 42:69-80
14. Wang, Weijie & Lu, Yanmin. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. IOP Conference Series: Materials Science and Engineering. 324. 012049. 10.1088/1757-899X/324/1/012049.
15. B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan and V. Vijayaraghavan, "A machine learning approach for prediction of on-time performance of flights," *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, St. Petersburg, FL, 2017, pp. 1-6. doi: 10.1109/DASC.2017.8102138

AUTHORS PROFILE

Varun Iyer is a final year Computer Science Undergraduate student at SRM Institute of Science & Technology, Chennai , India.

Karan Matalia is a final year Computer Science Undergraduate student at SRM Institute of Science & Technology, Chennai , India.

Dr.K Sreekumar is an Assitant Professoe in the department of Computer Science & Engineering at SRM Institute of Science & Technology, Chennai , India.