Machine Learning Approach to Flight Delay Prediction

Amy Edwards

March 19, 2019

Abstract

In 2015, the United States saw more than 5.8 million domestic flights [1]. Commercial aviation relies on a myriad of moving parts, which leaves plenty of room for error, resulting in a delayed or cancelled flight. This inconvenience is costly to travellers, airlines and airports. In this paper, we first look at a literature review of approaches to building flight delay prediction models. We then present a machine learning classification model that predicts whether a flight will have a significant delay (60 minutes or more) or be cancelled. Using random forest classification and flight information (origin airport, destination airport, airline, scheduled departure, and delay magnitude), it was possible to predict whether a flight will be delayed or not 64% of the time. Future work could increase this prediction by taking into account the weather, seasonality, and use parallel computing schemas.

1 Introduction

A flight is considered delayed by the Federal Aviation Administration (FAA) if it arrives or departs the gate 15 minutes or more after the scheduled time [2]. It is considered cancelled if the flight does not operate. Airlines report flight delays as one of five broad types: Air Carrier, Extreme Weather, National Aviation System (NAS), Late-Arriving Aircraft, or Security [1]. In 2015, nearly 20% of domestic flights in the United States were delayed or cancelled [3]. Delays are usually caused by unpredictable circumstances [10], and are therefore, extremely difficult to predict. In 2015, Air Carrier Delay, where it is the airline's fault, was the leading reason for delays [1]. Only when the delay is the airlines' fault are passengers entitled to compensation by the airline such as refreshments, rerouting, or a free room if the next flight is not until the day after. Passengers may be able to get help from the airline in other cases, but it is not required. Delays cost passengers if they are not due compensation from the airline. Passengers also have to plan ahead just incase of delays – possibly by travelling the day before an appointment, which results in increasing trip costs [4].

Delays cost the airline because in addition to the times they compensate passengers, they can be fined by the Transportation Department. This fine is for planes with passengers left on the tarmac for more than 3 hours without taking off on domestic flights. Airlines could see a decrease in profits if they have frequent delays because their reputation for punctuality is a key factor when people make the decision to book. In total, the FAA estimates that delays cost airlines $22 billion yearly [3] while the total cost to the US economy was estimated to be $32.9 billion in 2007 [2]. This cost does not factor in the environmental damage that is done by increasing fuel consumption and gas emissions from waiting planes or increased travel for the consumer [4].

Predicting or estimating flight delays can improve the tactual and operational decisions of airports and airlines. If the institutions are more knowledgeable, they can inform passengers of delays ahead of time, so they can make adequate plans [4]. Considering airlines already insert buffer into flight scheduling, quantifying even the propagated and newly formed delays would be an economic success [5]. This problem is made more challenging by the massive amount of data available for every moment of commercial aviation. In order to extract useful information from this vast volume, data scientists turn to machine learning algorithms for automatic (iterative) analytical model building.

The rest of the paper is structured as follows. Section 2 is a review of data and methods that have been used to predict flight delays using machine learning. Section 3 explains the methods used to obtain a predictive model for flight delays. Sections 4 and 5 summarize the results and discuss what they mean. Sections 6 and 7 discuss what future work can be done on the subject and present my conclusions.

2 Literature Review

There has been a lot of research on predicting and analyzing flight delays and their causes. Flight delay puts economic strain on the entire system. They play a vital role in air traffic control, airline decision-making [11], and passenger satisfaction. Various groups have looked at classifying the delays and performing regression analysis to determine the length of the delay. Models have been done with a variety of features – including specific airport pairs and incorporating weather at the airport.

In 2014, Rebollo and Balakrishnan presented a class of temporal and spatial models to predict departure delays 2-24 h into the future [6]. They consider classification models to determine if there will be a delay or not, and then regression models where the continuous output is an estimate of the departure delay time. They took into account high, medium, and low delay days between the main delay centers – airports in New York City, Atlanta, and Chicago. Their Random Forest model was trained on the 100 most delayed airport pairs. When classifying delays as above or below 60 minutes, they showed a test error of 19% for a 2 h forecast horizon and 27.4% error when increased to 24 hours.

Khanmohammadi et al. proposed a novel artificial neural network (ANN) using multi-level input as a prediction model for forecasting the magnitude of airport delays [7]. The new structure handles nominal variables better than a regular ANN. They trained with inbound flight data from JFK airport. They were able to get an RMSE of 0.1366, while a more common method of Gradient Descent Backpropogation had a higher RMSE of 0.1603. This shows an improvement with the new ANN method, yet only is developed for a single airport and further study will need to be done to see if it scales well.

In 2016, Belcastro et al. used flight information and weather conditions from January 2009 to December 2013 to predict arrival delay using the Random Forest data classification algorithm implemented with MapReduce and executed on a cloud infrastructure. They achieved 74.2% accuracy for a threshold of 15 minutes and 85.8% accuracy for a threshold of 60 minutes [8]. Using the Cloud allows for scalability as data available increases in the future. Without considering weather conditions, they also achieved an accuracy of 69.1%, which confirms that there is a pattern present concerning flight delay.

Choi et al. found that including the weather models improved predictive ability of their flight delay model [9]. Just like Belcastro [8], they chose to focus on arrival delays. They used US domestic flight data and weather data from 2005 to 2015. They applied sampling techniques to account for the imbalanced data and tuned a variety of machine learning techniques, where the Random Forest model was found to have the highest accuracy. Their most significant take away was that adding the weather data improved the accuracy across all models. The model may not perform as well when considering types of delays other than weather.

Manna et al. were the first to use Gradient Boosted Decision Tree as their machine-learning model to predict air traffic delay in 2017 [10]. They separated arrival delay and departure delay even though the two were highly correlated. Their model for arrival delays accounts for 92.3 % of the data variability and their model for departure delays accounts for 84.9% of variability. However, this model was only tested on the 70 busiest airports between April and October 2013. It may not be able to scale well.

To conclude the literature review I want to look at a two stage predictive model using flight and weather features done by Thiagarajan et al. They used Gradient Boosting to classify if a delay would be present (average 90.4% accuracy) and then Extra-Trees Regressor to find the magnitude of the delay (average 48.26 MSE) [11]. One key way they achieved this low error in regression was by using selective training. They trained the model on single airport-destination pairs individually, similar to Rebello [6], and then put it together in the final model.

3 Methodology

3.1 Preprocessing

The goal of this project is to find a generalized model for predicting if there will be a flight delay on any day of the year, given only the statistics known about the flight itself. The data used was downloaded from Kaggle.com, originally of the Bureau of Transportation Statistics (BTS) [1]. It covers all of the year 2015. I used 20% of the 5.8 million data points for creating the model because I was limited by computing power.

There were 30 columns of features included with the original data: *year*, *month*, *day*, *day-of-week* (1-7 corresponding to Monday – Sunday), *airline*, *flight number*, *tail number* (aircraft identifier), *origin airport*, *destination airport*, *scheduled departure*, *departure time* (wheels off – taxi out), *departure delay* (total delay on departure), *taxi out* (the time duration elapsed between departure from the origin airport gate and wheels off), *wheels off*, *scheduled time* (planned time amount needed for the flight trip), *elapsed time* (air time + taxi in + taxi out), *air time* (wheels on – wheels off), *distance* (distance between two airports), *wheels on* (aircraft wheels touch the ground), *taxi in* (the time duration elapsed between wheels-on and gate arrival at the destination airport), *scheduled arrival* (planned arrival time), *arrival time* (wheels on + taxi in), *arrival delay* (arrival time –scheduled arrival), *diverted* (aircraft landed on airport not in schedule), *cancelled*, *cancellation reason* (a - airline/carrier; b - weather; c - national air system; d – security), *air system delay*, *security delay*, *airline delay*, *late aircraft delay*, and *weather delay*.

Not all of these variables were needed and the rest needed to undergo preprocessing. I began with the airports because some of the airports were listed using five-digit numbers instead of three-letter code. I used code by Scott Cole as a template for converting everything to the three-letter code[12]. The airport code data for transformation was taken from the BTS [1].

I removed year, because the dataset was all from 2015. Day, flight number, tail number were removed because I was focusing on generalizing. Many features were combinations of others so I removed taxi out, wheels off, scheduled time, elapsed time, air time, wheels on, taxi in, scheduled arrival, and arrival time for this reason. I removed cancellation reason and the various delay reasons because I was simplifying to only look for if there was a delay or not. I created a new feature, 'late or cancelled', to be the target variable for classification. The FAA defines delay as being > 15 minutes, but has no definition of a "significant delay". I decided that a delay greater than one hour would be a significant inconvenience for people, so that is where I chose to make the threshold. The target variable becomes any flight that was delayed more than an hour or cancelled altogether (Fig. 3).

Time of departure is a categorical feature, so I divided it into 4 sections to make sense of the time of day: 00:00 – 06:00 hours is early morning, 06:01 – 12:00 hours is late morning, 12:01 – 18:00 is afternoon + early evening, 18:01 – 24:00 is evening (Fig. 1). Airports that contributed to less than 5% of the total airports were deleted, in order to focus on the bulk of the data. This removed airports that had less than 3857 incoming or outgoing flights and left the top 58 airports (Fig. 2).

To explore the data more, I plotted the frequency of our target variable – late or cancelled – against the month of the year, day of the week, airline, and scheduled departure time. February, June, July and December are the months with the highest delays while the shortest delays are found in September and October. Monday has the most delays and surprisingly, Saturday has the least delays. Predictably, the earlier the flight, the less likely there is to be a delay. Proportional to the number of flights, it shows that MQ, Envoy Airlines, has the most delays or cancellations, while HA, Hawaiian Airlines, has the least delays (Fig. 1).

In order to parse more information from the air-

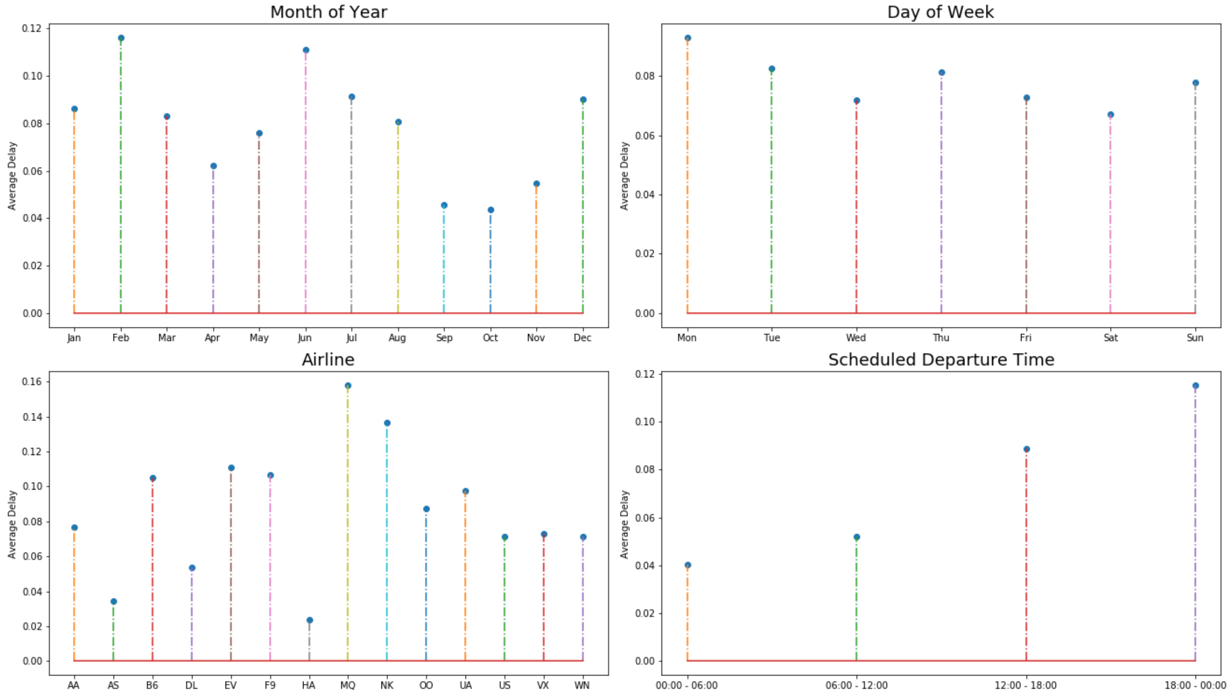

Figure 1: Various exploratory graphs

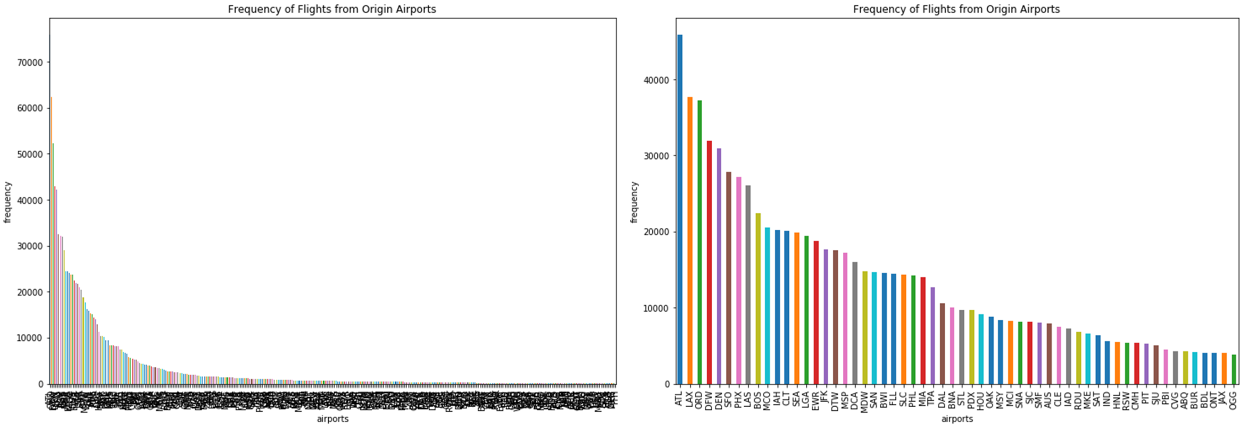

Figure 2: Origin airports before and after reduction

lines, I created a factor called "airline risk" based on the proportion of delays per the number of flights that each airline has. I then removed the airline categorical variable. This actually ended up decreasing the accuracy, but I think the idea is a good one. I will explore it further in future work.

Finally, the categorical variables were converted to dummy variables to be used in classification. Once done with data cleaning, I saved 10% of the data for validation after training and testing the model. In order to account for the disparity between flights that are on time and flights that have been significantly delayed or cancelled, I used an under sampling method on the other 90% of the data for training (Fig. 4). The training data set ended up at 33.9 MB and 149 columns. My features were distance, diverted, month, day of the week, airline, origin airport, destination airport, and scheduled departure. The target was the created feature "late or cancelled". The bulk of the columns were dummy variables because distance was the only continuous variable.

3.2 Predictive Model

I tried a variety of machine learning methods: Decision Tree, Random Forest, Gradient Boost, Ada Boosting and Neural Networks. To evaluate the machine learning methods, I used a 35% train/ 65% test split and recorded accuracy and AUC. The method with the best scores was Random Forest. I used a grid search algorithm to find the best model parameters. The random forest model was best with 75 estimators, 20 maximum features, a max depth of 30, trained on entropy and with a minimum sample split of 3. Setting the number of features to consider at each stop did not seem to have that much effect on the model accuracy. We looked three models with all of the same hyper-parameters expect for the class weight method. The three models were: 1. A random forest model without a weighting scheme, 2. A random forest model weighing the positive delay class more, and 3. A random forest model weighing the "no delay" class more.

To select features, I tried a combination of brute force, Stepwise Backwards Selection, partial Wrapper, Univariate Chi-Squared, ExtraTrees, and PCA.

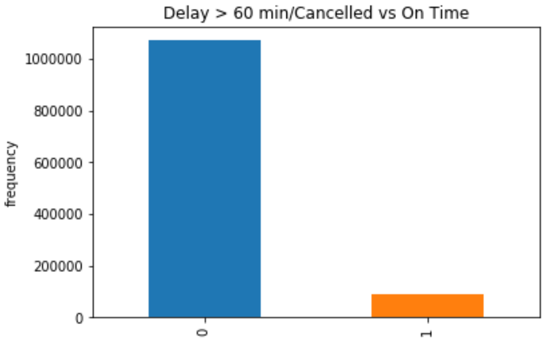

Figure 3: data before undersampling

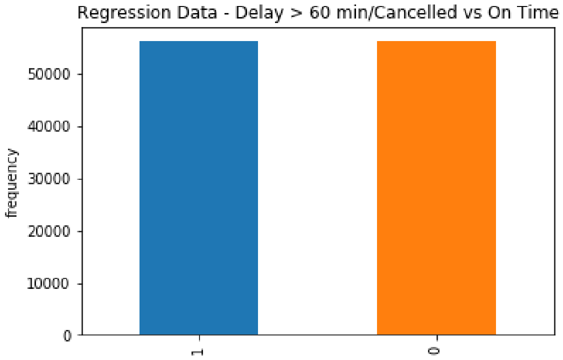

Figure 4: data after undersampling

I looked for collinearity using a heatmap and printed the variance inflation factor (VIF) values. None of the VIF values were over 10, which is where there would be concern and the heatmap did not show collinearity that was concerning. Removing any of the features or using principle components resulted in a dramatic loss of accuracy, so I decided to leave them all in.

4 Results

To evaluate the results, I looked at accuracy, AUC, the confusion matrix (True Negatives, True Positives, False Negatives, and False Positives), Precision, Recall, Specificity, and the F1 measure. The results shown here are from the training set, testing set, and then the 10% holdout validation set. The target has been balanced in training and testing set, but not for the validation set (See Table 1).

Accuracy is the ratio of correctly predicted instances to all the instances. It can be skewed when looking at unbalanced data. In the case of flight delays, it is much more likely for the flight to be on time than it is to be delayed, so a random guess of "no delay" can achieve 100% accuracy. The final unweighted model here has an accuracy of 64%. This is similar to what others have got using only flight statistics [8].

$Accuracy = TP + TN/(TP + FP + FN + TN$

AUC stands for "area under curve". The curve is Receiver Operating Characteristic curve, which shows the trade off between the ratio between the true positives rate and the false positive rate. AUC tells how much the model is capable of distinguishing between classes, or "separability". The higher the AUC, the better the model can distinguish between flights that are delayed and flights that are on time. A random chance AUC = 50%. The unweighted model generated here had the best overall scores and achieved AUC = 61.7%, which is slightly better than random chance.

Precision is the ratio of true positives to all predicted positives while recall is the true positive rate, also called sensitivity. They show how well the model correctly picks out the delayed flights.

$Precision = TP/(TP + FP)$

$Recall = TP/(TP + FN)$

Specificity is also called the true negative rate. In relation to this dataset, it is the proportion of on time flights that are correctly identified as on time. This is particularly important because it is better to mistakenly predict that there will be a delay than to mistakenly predict that there will not be a delay, when there is.

$Specificity = TN/(TN + FP)$

For binary classification, the F1-Score is a measure of accuracy that considers both precision and recall where 1 is the best score. Because the validation set has a target true to life – extremely unbalanced – it is important to consider the F1-Score rather than accuracy.

$F1 - score = 2 * (Recall * Precision)/(Recall + Precision)$

The best model was without weighting either of the classes, due to the convoluted ways the weight changed the confusion matrix. When correctly predicting that there was not a delay mattered more, the model essentially guessed that there would not be a delay. When predicting that there would be a delay was weighted more, the model had mostly false positives. The unweighted model has a balance between recall and specificity as well as the highest F1-Score of 0.22.

5 Discussion

The goal of this paper was to develop a general predictive framework in attempt to address a crucial issue in aviation – delays on incoming and outgoing flights that create economic stress, flight management issues, and decrease an airline's reputation. This model was intentionally extremely general – only taking into the statistics known about a current plane flight and when it takes place. The results show that the approach is worth further consideration and that it is entirely feasible to predict a future delay. The model currently only performing binary classification, but can be extended to regression in order to predict the magnitude of the delay. It can potentially be useful to airports for management of the days' flights, to airlines to know how to best prepare, and to customers to aid them in time management and financial planning.

6 Future Work

In the future, using parallel machine learning algorithms on scalable computing systems, as others have done [8] would be valuable for creating a more general model. I think it would be good to explore adding a regression model after classifying like Thiagarajan et al. did[11] in order to get the scope of the delay. A delay of three hours is significantly different from a delay of half an hour. I looked into this briefly, but it would take a very significant analysis on the outliers in delay time. Since most flights are on time, this regression would really be trying to predict the

	class weight = None			class weight = {0:0.9, 1:0.5}			class weight = {0:0.5, 1:0.9}		
	Train	Test	Validate	Train	Test	Validate	Train	Test	Validate
Accuracy	82%	64%	64%	54%	54%	90%	56%	55%	55%
AUC	92%	69%	61.7%	69%	67%	53.9%	69%	67%	21.6%
TN	28,906	12,073	44,858	35,494	19,137	70,857	5,894	2,952	11,033
FP	7,403	7,604	28,373	866	579	2,273	30,376	16,764	62,097
TN	5,386	6,656	2,114	17,316	17,316	5,677	1,337	846	266
TP	31,086	12,858	4,258	2,159	2,159	695	35,174	18,629	6,106
Precision	0.81	0.63	0.13	0.71	0.79	0.23	0.54	0.53	0.09
Recall	0.85	0.66	0.67	0.11	0.11	0.11	0.96	0.96	0.96
Specificity	0.80	0.61	0.61	0.98	0.97	0.97	0.16	0.15	0.15
F1	0.83	0.64	0.11	0.19	0.19	0.15	0.69	0.68	0.16

Table 1: Results

magnitude of the outliers.

A more robust model could also come by adding more features. Adding weather seems to improve the delay predictions significantly [6, 8, 9], so that would be something to explore. These groups only included the weather conditions at the origin or destination airports. I would like to take into account the weather on the flight path. In addition, it might be useful to take holidays into consideration. Accuracy may be improved if separate models are developed for different times of the year, rather than the general model that was presented in this paper. Looking into the delay risk associated with different airports and using that value instead of the airport is another avenue to explore [6].

7 Conclusions

In this paper, a bivariate classification model for predicting flight delay was presented. On the massively unbalanced dataset, it predicted delay or no delay correctly 64% of the time. This is consistent with results obtained from other research [4, 5, 6]. The fact that there has been so much research already done on measuring and predicting flight delays shows just how critical the subject is. Adding weather features and additional years of data can easily extend the model and increase the accuracy. In practice, this model can be used for airlines and airports to know where they need extra resources. It could assist them in combating the delay as well, because the ideal scenario is that there are not any delays. The output could be made available online for customers to have a picture of what delays are happening, rather than waiting for last minute updates. Since predicting flight delay accurately is crucial to a well functioning aviation system, this topic is worthy of further study.

References

[1] Department of Transportation. “2015 Flight Delays and Cancellations.” Kaggle, 9 Feb. 2017, www.kaggle.com/usdot/flight-delays/home.

[2] “OST_R | BTS | Transtats.” BTS, transtats.bts.gov/.

[3] Rapajic, Jasenka. Beyond Airline Disruptions: Thinking and Managing Anew. Routledge, 2018.

[4] Sternberg, Alice, et al. “A Review on Flight Delay Prediction.” ArXiv.org, 3 Nov. 2017, arxiv.org/abs/1703.06118.

[5] Kafle, Nabin, and Bo Zou. “Modeling Flight Delay Propagation: A New Analytical-Econometric Approach.” Transportation Research Part B: Methodological, vol. 93, 2016, pp. 520–542., doi:10.1016/j.trb.2016.08.012.

[6] Rebollo, Juan Jose, and Hamsa Balakrishnan. “Characterization and Prediction of Air Traffic Delays.” Transportation Research Part C: Emerging Technologies, vol. 44, 2014, pp. 231–241., doi:10.1016/j.trc.2014.04.007.

[7] S. Khanmohammadi, S. Tutun, and Y. Kucuk. A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. In Procedia Computer Science, volume 95, pages 237–244, 2016.

[8] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. "Using Scalable Data Mining for Predicting Flight Delays: A General-Purpose Network Analysis and Graph-Mining Library." Acm Transactions on Intelligent Systems and Technology (tist) 8, no. 1 (2016): 1-20. doi:10.1145/2888402.

[9] Choi, Sun, et al. “Prediction of Weather-Induced Airline Delays Based on Machine Learning Algorithms.” 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016, doi:10.1109/dasc.2016.7777956.

[10] Suvojit Manna, Sanket Biswas, Riyanka Kundu Somnath Rakshit, Priti Gupta and Subhas Barman .“A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree”.2017 International Conference on Computational Intelligence in Data Science(ICCIDS). 978-1-5090-5595-1/17/$31.00 ©2017 IEEE.

[11] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, Vineeth Vijayaraghavan[4] . “A Machine Learning Approach for Prediction of On-time Performance of Flights “.978-1-5386-0365-9/17/$31.00 ©2017 IEEE.

[12] Cole, Scott. “Fix Inconsistent Airport Codes.” Kaggle, 2017, www.kaggle.com/srcole/fix-inconsistent-airport-codes.