Flight Disruption Insights with Big Data Analytics

1st Krishna Bathula
Seidenberg School of CSIS, Pace University
Pleasantville,NY
kbathula@pace.edu

2nd Kaleemunnisa LNU
Seidenberg School of CSIS, Pace University
Pleasantville,NY
klnu@pace.edu

3rd Dr. Charles C. Tappert
Seidenberg School of CSIS, Pace University
Pleasantville,NY
ctappert@pace.edu

4th Istvan Barabasi
Seidenberg School of CSIS, Pace University
Pleasantville,NY
ibarabasi@pace.edu

5th Dr. Ronald I. Frank
Seidenberg School of CSIS Pace University
Pleasantville,NY
rfrank@pace.edu

***Abstract*—Data is exponentially growing and in different forms such as structured, unstructured and semi-structured. This study uses big data analytics concepts and machine learning algorithms to work with structured data of flights, airports, airlines and weather information. The objective is to represent the correlation between different data points among the datasets and use these associations to identify the key features that can disrupt flight schedules and lead the study for impact analysis. The domino effect that is passed on to the stopover and connecting flights in the route to their destinations is also predicted. These insights provide the basis for disaster management and recovery of valuable air time as the delays in flights influence economy of airport authority, airlines and flyers, causing damage to environments due to increase consumption of utilities like fuel and gas.**

***Index Terms*—Big data analytics, data science, data mining, flight schedule disruptions, flight delays, weather analysis.**

I. Introduction

The U.S. Department of Transportation tracks the punctuality of domestic flight arrivals operated across the USA by large airlines. This statistical information has an apparent increase of air transportation in recent years, and points towards progressively congested airports and airspaces. This extensive usage of airspace reflects the heightened risk of operational disruptions that can be traced to delays, cost criteria, deteriorated quality of service, airline setbacks, passenger discontent, etc. In addition to these, airlines are constantly finding numerous strategies in optimizing their operational profits above competing with their counterparts. It is a challenging task to balance higher turnover with customer gratification and sustainability [1].

As per the analysis done by Bureau of Transportation Statistics (BTS), nearly 20 percent of commercial air transport run out of scheduled time. Which results in heavy losses to airline companies as well as superfluous distress to customers. Weather is contributing sophisticated in flights delay as well as late arrivals. The study of prediction of delay in flights is vivacious area of research as demands for air travel increase. according to the U.S. Department of Transportation's Air Travel Consumer Report, in January 2017 alone, all the airlines of United States of America, recorded a mean on-time arrival frequency of 76.0 percentage, down from 81.0 percentage compared to 2016 data and up to 82.5 percentage in the final quarter of 2016. Thus, various studies were attempted earlier to determine patterns in air traffic and flight schedule deviation from the actual planned.

In this study, we focus on the flight disruptions that arise from air system delays, security delays, airline delays, late aircraft delays and weather delays. Our aim is to find the correlation between attributes like flights and weather and whether we must consider all the constraints that effect delays or just the ones that have major impact. This impact analysis further provides grounds for determining the influence of these delays to the connecting flights from the destinations. The data from various datasets are observed for the features that effect the overall delays. These patterns in the itinerary and delays can be used to predict the possible future delays. We use data science algorithms to do in-depth study of several key factors that contribute to the airline performance factor and determine the correlation among these characteristics. We also propose a model that will predict the flight disruptive phenomenon with respect to climate changes and forecast the repercussions to the consecutive flight patterns, one such instance being the connecting flights. These observations will help the airline industry to take essential precautions for operational effectiveness, time and wealth optimizations.

II. Literature Review

According to the latest reports by the Bureau of Transportation Statistics, U.S. Department of Transportation [2], the on-time flights are 78.37 percent. The rest of 19.68 percent constitutes to the delay in schedule. A flight can be announced delayed if it arrives more than 15minutes than planned schedule. Although the various reasons for delay like weather, security, National Aviation System (NAS) delay etc., are considered, if there are multiple reasons contribute to a

delay, each source is prorated based on the deferred minutes. The results for on-time arrival are analyzed from the source of the historical data from June 2003 to August 2019 (Fig. 1).

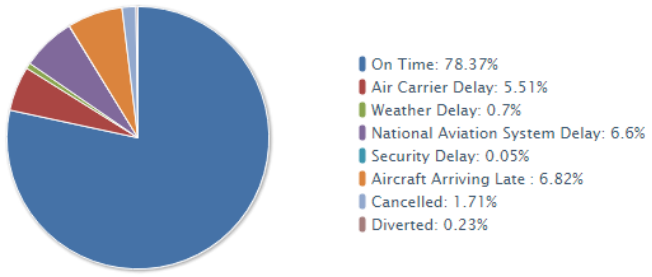

Fig. 1. U.S. Department of Transportation Statistics.

If the historical data is further prompted for the exact cause of delay, the study yields weather as the much incriminating factor with almost 61.89 percentage than the others (Fig. 2). All these measures are based on the number of flight operations.

Causes of National Aviation System Delays
National (June, 2003 - August, 2019)

	Number of Operations	% of Total Operations	Delayed Minutes	% of Total Delayed Minutes
Weather	4,384,592	61.89%	237,732,919	72.99%
Volume	1,831,716	25.86%	57,336,563	17.60%
Equipment	51,160	0.72%	2,324,560	0.71%
Closed Runway	564,633	7.97%	19,546,609	6.00%
Other	252,329	3.56%	8,786,129	2.70%
Total Operations	7,084,430	100.00%	325,726,780	100.00%

Fig. 2. Flight Delay Analysis Report (Source BTS DOT, US).

Rebello et al. have created indicative model predicting network-related interruptions of the forthcoming by applying the system-level dependences between airports [6]. In a similar fashion, Hansen et al. analyzed the advancement in flying delays in the United States, domestic system by assessing an econometric model of systematic routine delay that combines the properties of arrival line up such as terminal weather situations, seasonal and secular impacts. The results suggested that even after monitoring these factors overall, the delays decreased gradually from 2000 through mid-2003, but the trend inverted drastically thereafter [4].

Another group of researchers Mueller et al, developed a statistical method to analyze the departure, arrival data and characterize the delay data [5]. Belcastro et al. [2] developed a model that govern onset delay prediction of a scheduled flight. On the other hand, Choi et al. [3] have anticipated a model to forecast carrier onset interruptions caused by inclement weather situations using data mining techniques and supervised machine learning algorithms.

Over a period, numerous analytical models and simulation methods have been developed to analyze flight delay, which includes deterministic queuing models, neural networks, econometric models etc.

III. Methodology

In this research we are predicting the flight delay due to weather disruption, which can help airlines and passengers to have appropriate plan of action. In this paper we will be using machine learning techniques for predicting the delay. The project is divided in to 3 different parts, one is the Data Engineering, the second is the Exploratory Data Analysis and the last is the prediction of the connecting Flight Disruptions.

The initial step towards a successful Data Analysis is ensuring the Data Quality. The Airline data is fetched from Department of Transportation US (DOT). And the source of Weather data is National Oceanic and Atmospheric Administration (NOAA).

A. Data Preprocessing

Data is thoroughly examined for integrity criteria as well. Since we expect the model to work with all the forms like offline, near line and online data, we curtailed the irrelevant and unnecessary parameters that could overburden the dataset. We have also dropped the null values and assigned zero to Not a Number (NaN) values as one of the data cleansing activities. The data types of time factors such as scheduled time, airtime etc., are found to be in float point and needs proper conversion of input time to standard date time format. The categorical data is assigned with proper numeric values which are the most contributing factors that flag the key filters. Finally, the data is analyzed for distribution after cleansing, converting and preprocessing.

Then different datasets such the airline, flight, airports and weather datasets are integrated and normalized to identify the correlating factors that affect the flight cancellations (Fig. 2).

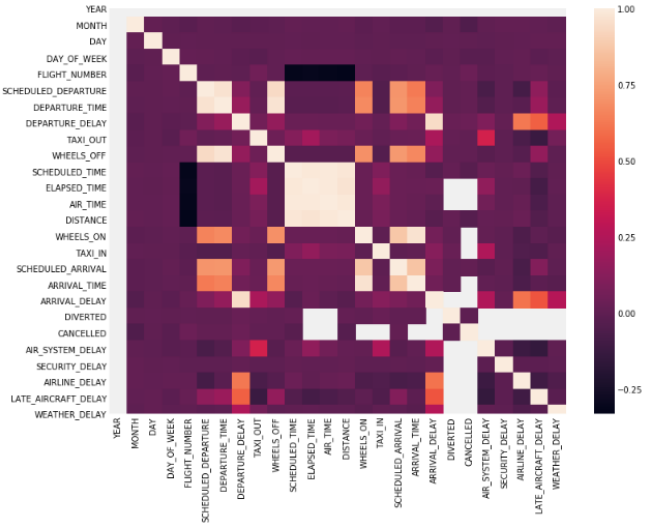

Fig. 3. Feature Correlation

B. Exploratory Data Analysis

As part of EDA, we have explored data in different views. First EDA is on Number of flights per year at all airports. The distribution of delays by airlines is reasonable analysis as seen in pair plot Fig. 3, with respect to the scheduled time, geographical area covered, and the delay incurred. Hence, we will be considering the airline characteristics as key features for our model.

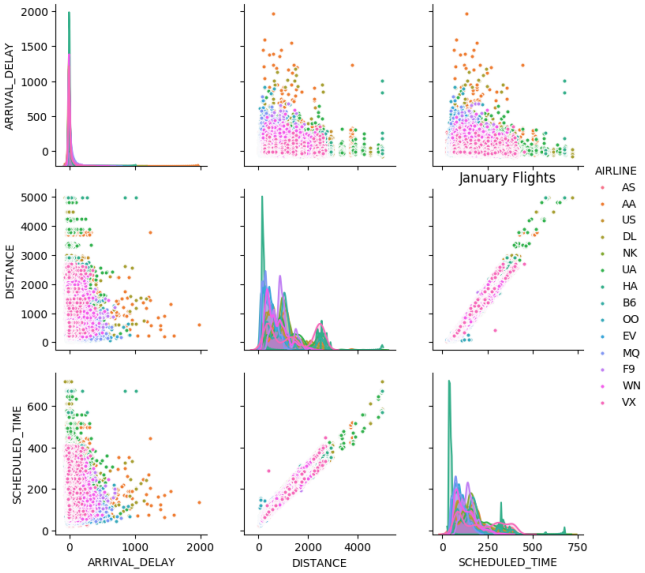

Fig. 4. Pair Plot

Since all the airlines are operated from airports which can be either an origin or destination. From these data we will be selecting only Top 50 busiest airports. To get the Top 50 we have added departures and arrivals. The predictors are chosen based on their delay factors. The data frame for the small feature set is fitted with Random Forest classifier and extracted Feature importance score for each feature. Such as Departure Delay and Arrival Delay due to weather.

1) Departure Delay Prediction: All the feature set utilized in the departure delay forecasting is identified with the help of correlation matrix. Therefore, in predicting the delay, each feature is relatively significant.

2) Arrival Delay Prediction: As departure delays will affect the arrivals, we will be selecting the origin airport.

Data Integration: Top 50 busiest airlines data is integrated to weather data by origin and destination airports at the time of takeoff and landing. While analyzing the data we have found that approximately 65 percent of the flights originate and land in these airports. The Flight data is integrated with the weather data, for all the weather station, we have considered the average weather parameters, i.e. Annual Mean Temperature, Annual Mean Precipitation, Annual Mean Visibility etc. Two Data Frames are created for simplicity - One for Origin, and one for Destination. They are the same dataframes, except for the Column Names.

The delays are plotted against weather conditions such as heavy rains gauged in inches fig.5. The impact analysis plot can be viewed in fig.4

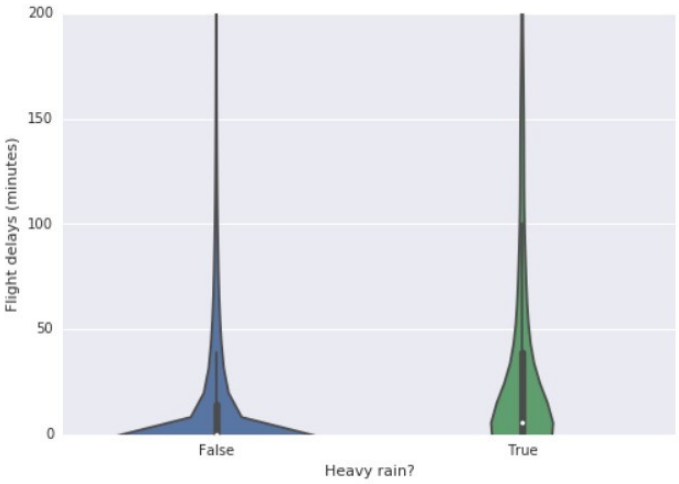

Fig. 5. Rain effected Delays

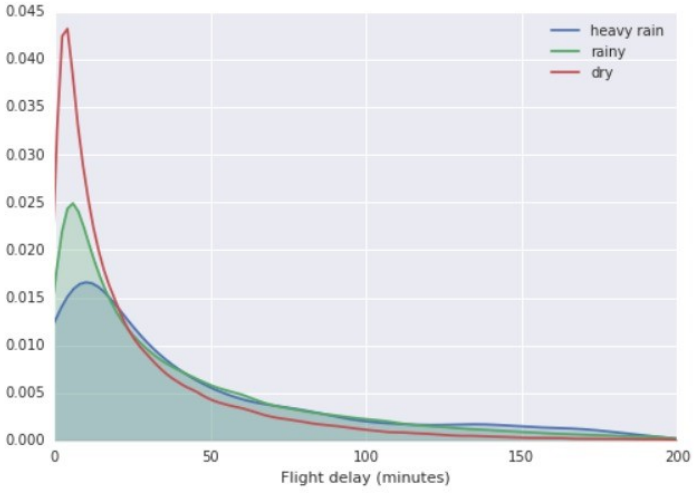

Fig. 6. Delays vs Precipitation Rate

C. Model

The input data is split in to training and testing data. The intent of our model is to predict arrival delay, which gives us the reference window time prediction for the connecting flight. In order to get closer time window that determines if a connecting flight can be boarded by delayed passengers, we start with arrival delay which is vastly tricky, as majority of flights having zero or a small arrival delay. We break the problem into two sub-parts:

1) Delay Classification Model

The threshold of delay factor being more than 5 minutes, we performed binary classification, training a logistic regression model and record resulting P value of the delay i.e., the output probability of delays.

2) Predicted Delay

Perform Linear Regression and model trained on positive delays from the above result of binary classification.

D. Model Evaluation

The model has been trained to predict arrival delays, given flight features such as flight number, origin and destination etc. Additionally, the weather features like precipitation, wind speed, visibility which are primarily key characteristics. To discount the effect of weather on historical delays, we predict arrival delay for each flight with the mean weather conditions of the origin and destination airports.

IV. Results

As per the preliminary results, we were able to obtain 30 percent of positive prediction Fig. 6, which is on the much lower side. We have selected the K-fold cross validation method which is deterministic. If we try to use more than one model, there is a possibility of over fitting the data and this approach may lead the parameters of the model to be biased. We were able to achieve these results using the regularization techniques.

```
Final Accuracy: 0.712627
TP: 158341
FP: 145880
TN: 554286
FN: 141493
% of positive predictions:
0.304221
```

Fig. 7. Results

V. Conclusions

With the help of this model airlines and passengers can get advance notice on the expected delays in their journey. This study proposed a prediction model which is Regression using Linear regression and Logistics Regression Model classify the delay. Airline delays triggered by extreme weather condition. In specific, the model was built on historic weather and airlines data for top 50 airlines by utilizing machine learning algorithms. The project has incorporated and showed the importance of Regression Analysis in Machine Learning, Big Data Analytics, and also Cross Validation technique and Regularization in ML for making proper models. Because the data was imbalanced, we have performed data cleansing techniques. We were also able to infer that there is significant subsequent impact on the connecting flights when there is a delay of greater than 45 minutes in the arrival of the aircraft at the stopover destination.

The model's prediction performance on the validation set and the test set was analyzed. We feel that there are few more possible methods that can be useful to improve the model in the future. Also, for future work we can integrate expenses which could be saved by predicting the delays, every second in delay could lead to losses. We will study on how we can save one percent by avoiding delays which are in control.

References

[1] Samet Ayhan, Johnathan Pesce, Paul Comitz, D. Sweet, Steve Bliesner, and Gary Gerberick. Predictive analytics with aviation big data. pages 1–13, 04 2013.

[2] L. Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology*, 8, 01 2016.

[3] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. pages 1–6, 09 2016.

[4] Mark Hansen and Chieh-Yu Hsiao. Going south?: Econometric analysis of u.s. airline flight delays from 2000 to 2004. *Transportation Research Record*, 1915:85–94, 01 2005.

[5] Eric Mueller and Gano Chatterji. Analysis of aircraft arrival and departure delay characteristics. 10 2002.

[6] Juan Rebollo and Hamsa Balakrishnan. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 44:231–241, 07 2014.