Analyzing relationship between Flight delays and Weather Data for different US airports

Aishwarya Budhkar
New York University
New York, United States

Varada Hanumante
New York University
New York, United States

***Abstract*—**

In this paper we study the relationship between flight departure delays and weather conditions for different US airports. The aviation areas generate huge amounts of data. Big data tools can be used in aviation areas to improve the performance and customer satisfaction. Historical and predicted weather data is available. Sensors are installed on aircrafts to monitor its performance and health. Weather, flight information and sensor data can be used to develop predictive model using big data analytics to enhance flight safety, improve customer satisfaction and reduce flight delays. Flight data from Bureau of Transportation Statistics and weather data from National Centers for Environmental Information, National Oceanic and Atmospheric Administration were used to analyze their relationship. Big data platform and tools were used for data processing and analytic.

Keywords—Big data applications, Machine learning algorithms, Analytical modeling, Data analysis, Data preprocessing, Data collection, Support vector machines, Decision Trees.

I. INTRODUCTION

Due to the huge growth in air traffic in recent years, it is required to develop effective air transport control systems. Data is collected from ground stations, satellites, sensors on aircraft leading to huge volume of data collected. GPS sensors collect information like distance, time of departure and arrival, etc. Weather stations record information about historical weather as well as predict weather conditions in future. The large volume of data collected from various sources is too big to handle for traditional systems [2].

As traditional systems find it difficult to process such huge data, big data framework is used to find patterns in data quickly. Big data is considered to have large volume, veracity and velocity. We find patterns, relations within data and other insights in big data analytics. Many frameworks are available for big data analytics. It is useful to prove or disprove assumptions and used on large scale due to fast development.

The remainder of the paper is organized as follows: Section II describes why the analytic is important; Section III includes the related research work; Section IV describes the datasets used for the analytic; Section V describes the insights derived from analytic; Section VI includes the design of application; Section VII informs about how the application is used and what action can be taken in response to the insight produced; Section VIII is discusses the experimental setup and limitations of the application; Section IX describes the results and use of the application; Section X states the possible improvements.

II. MOTIVATION

Airways is one of the important means of transport. In recent years due to easy of travel and low-cost flight availability air travel has seen a considerable increase. There is a need to develop effective management technique improving service to customers and giving efficient service to airline providers. According to reports, the annual cost due to air transportation delays was over 30 million dollars [5].

Flight delays are unavoidable in certain conditions. Predicting flight delays in advance will ensure that the customers are well aware of probable delay and adjust their schedules accordingly. This will save them time and money. Airline companies can also be made aware of any landing delays they may encounter they have enough resources or change their course if required. They can take necessary steps to accommodate passengers from cancelled flights in their upcoming flights and make arrangement for passengers at airport. Thus, they can increase revenues by customer satisfaction and ensure safety. Airport management can schedule landing and departures so as to accommodate delayed flights and ensure smooth and efficient operations.

III. RELATED WORK

There is a drastic growth in air traffic demand and effective management of operations [2]. A huge volume of data is extracted from sensors, ground stations and customer surveys. The vast data cannot be handled by traditional database systems. The various applications of big data analytics in aircraft infrastructure and operations is important to understand [2]. Several limitations like data manipulation, bias in data and security need to be addresses [2]. Flight delays case several loses like time, money and energy. Accurate delay prediction is crucial due to the increasing complexity of air transport [3]. Various factors can lead to delay like route delay, delay propagation, runway constrains, weather conditions, etc. Methods like probabilistic models, statistical analysis, machine learning and operational research can be used to predict the delays [3]. The methods commonly used in Machine learning are K-Nearest neighbors, SVM and random forests solving classification, prediction problems

[3]. Various methods have been used for flight delay prediction using Machine Learning, Deep Learning and big data. Big data technique used flight data as well as weather conditions and applies parallel algorithms using MapReduce for delay prediction based on weather conditions [4]. With arrival delay, propagation delay and departure delay affecting the delay of flights, a chained model can be used which will predict the total delay. Recursive feature selection and elimination is important [5].

Scheduled arriving aircraft demand may exceed airport arrival capacity when there is abnormal weather at an airport. In such situations, ground-delay programs are instituted to delay flights before they depart from their originating airports [7]. Efficient planning of such programs depends on the accuracy of prediction of airport capacity and demand in the presence of uncertainties in weather forecast [8]. Weather has various parameters like temperature, pressure, humidity and wind speed. Huge amount of data has been collected and archived in an unstructured format. Big data technology like Hadoop and Spark have evolved to solve the challenges and issues of big data using distributed computing [9]. Hadoop & Map Reduce are the most widely used models used today for Big Data processing. Apache Spark is the new competitor in the Big Data field [10]. Spark design has proved to be 100 times faster than Hadoop MapReduce in certain cases. Spark supports in-memory computing and performs much better on iterative algorithms, where the same code is executed multiple times and the output of one iteration is the input of the next one [10]. Study on the influence of inclement weather on airline delays is essential for efficient flight operations. A decision support tool based on the study can inform the passengers and airlines about weather-induced delays in advance and help them reduce possible monetary losses [11].

IV. DATASETS

We have used two datasets: Bureau of Transportation data and National Centers for Environmental Information, National Oceanic and Atmospheric Administration link data.

A. National Centers for Environmental Information, National Oceanic and Atmospheric Administration Link

Weather data was used from National Oceanic and Atmospheric Administration link [7]. Data for 2017 and 2018 was used. The data is static and collected once. The size is 0.5 GB. Of the many available fields in data following fields were used for the analytics:

Column Name	DataType (max_length)	Limits	Example Values
DATE	STRING (8)	1/1/17 - 12/31/18	8/23/17
AVG_WIND	DOUBLE (4)	0 - 29.97	13.2
PRECIPITATION	DOUBLE (3)	0-9.36	4.5
SNOW	INTEGER (3)	0-140	100
MAX_TEMP	INTEGER (2)	32-98	45
MIN_TEMP	INTEGER (2)	20-96	30

B. Bureau of Transportation – flight data

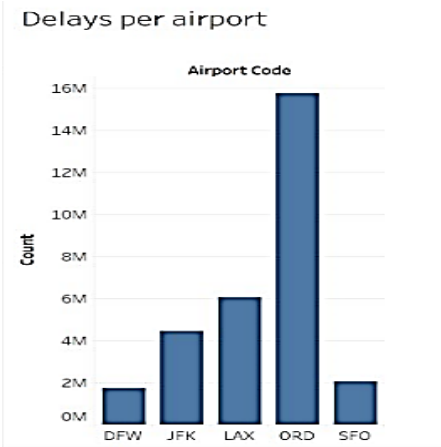
Flight data was used from Bureau of Transportation [1]. Data for 2017 and 2018 was collected. The data is static and collected once. It is 5 GB in size. Of the many available fields in data following fields are used for analytics:

Column Name	DataType (max_length)	Limits	Example Values
YEAR	INTEGER (4)	2017 – 2018	2017
MONTH	INTEGER (2)	1 – 12	4
DAY_OF_MONTH	INTEGER (2)	1 – 31	23
DAY_OF_WEEK	INTEGER (1)	1 – 7	2
ORIGIN	STRING (5)	272 different origins	“ABE”
DEST	STRING (5)	272 different destinations	“JFK”
CRS_DEP_TIME	STRING (6)	“0001”- “2359”	“0001”
DEP_TIME	STRING (6)	“0001”- “2400“	“2400”
DEP_DELAY	DOUBLE (5)	-84 .0– 2710.0	2710.0
ARR_TIME	STRING (6)	“0000” – “2400”	“2400”
ARR_DELAY	DOUBLE (5)	-104.0– 2692.0	2692.0
CANCELLED	DOUBLE (2)	0.0/1.0	1.0
DISTANCE	DOUBLE (5)	66.0 – 4983.0	1433.0

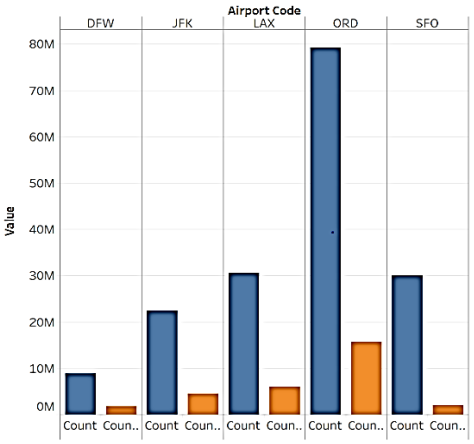
V. DESCRIPTION OF ANALYTIC

The relationship between flight and weather data was analyzed to figure out the factors affecting the flight delays with their importance. Data collected from Bureau of Transportation Statistics for 5 airports – JFK, ORD, SFO, LAX and DFW along with weather data collected from NOAA for these airports was analyzed.

The analysis was done at different levels gradually doing more granular study. Airport-wise data analysis was done. The results were:

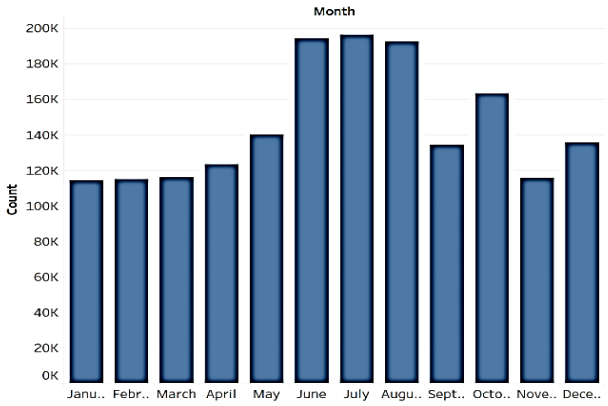


ORD airport showed the most departure delays followed by LAX, JFK, SFO and DFW.

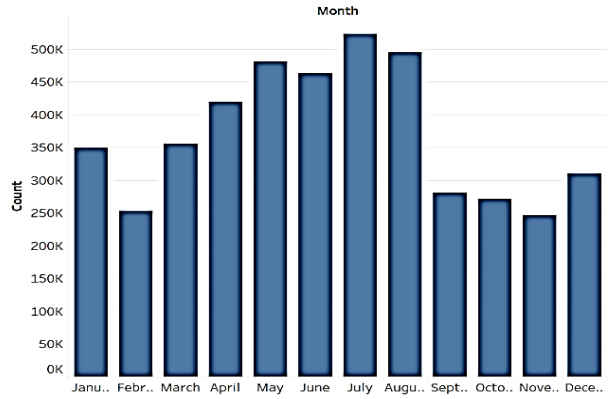


This shows that SFO has least delays which can strengthen that weather affects the flight delays but DFW, ORD, JFK and LAX show almost equal percent of delayed flights which leads to a conclusion that other factors need to be investigated apart from weather to predict the delays accurately. Month-wise data was analyzed for all the five airports. Results were as followed.

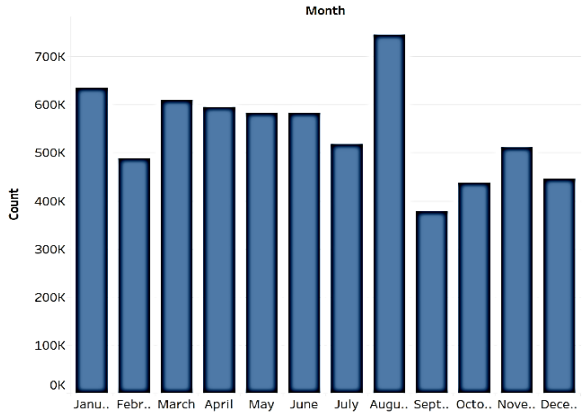
DFW monthwise delays



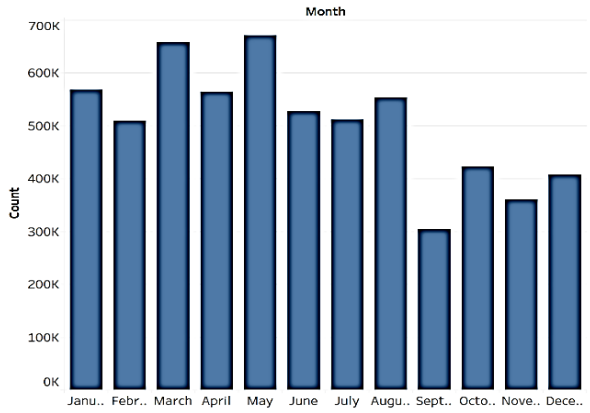
JFK monthwise delays



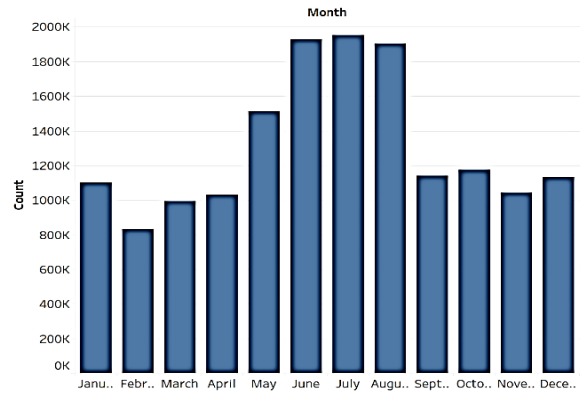
SFO monthwise delays



LAX monthwise delays



ORD monthwise delays



The flight departure delays were high at JFK from April to August. Similarly, in ORD and DFW showed high number of delays from June to August. While the delays distribution was spread evenly at LAX and SFO, less delays were observed from September to December.

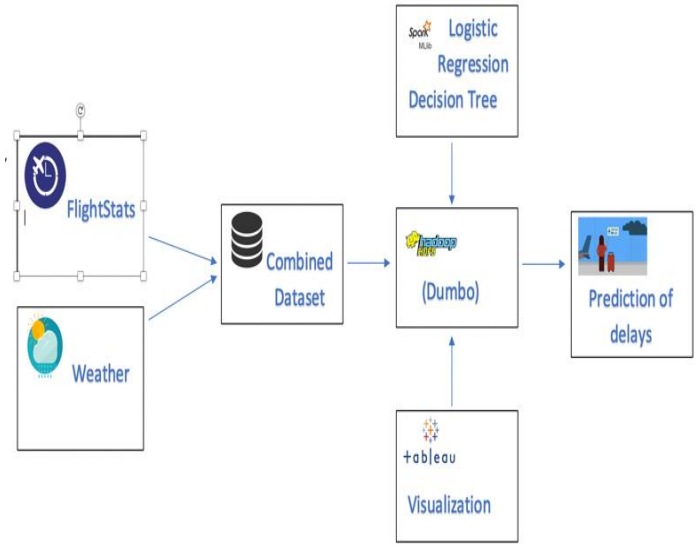
Distance wise departure delay distribution was checked monthly which showed a uniform distribution over different ranges of flight distance which leads to a conclusion that flight distance does not show significant effect on the delay.

As the data distribution was not uniform with non-delayed flight records exceeding four times the number of delayed records, Synthetic Minority Over-sampling Technique [13] was used to generate synthetic samples of minority class and thus make the data distribution uniform for machine learning algorithms.

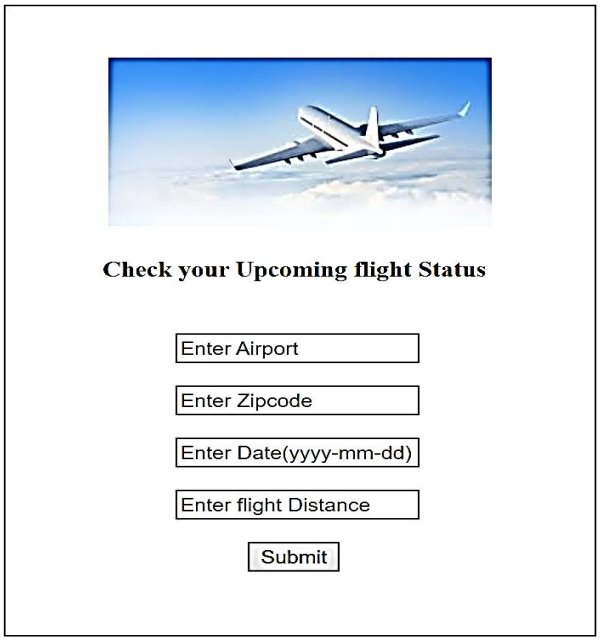
A machine learning model was trained to predict if a flight will be delayed using the weather data and flight information. Weather features – Precipitation, Snow, Wind speed, Max Temperature, Min Temperature and Flight features like arrival delay and flight distance were used to train the model. Logistic Regression, Support vector Machines, Decision Tree and Random Forest classifiers were used. 2017 data was used for training the model while 2018 data was used for testing. The binary classifier was trained for 100 iterations. Best testing accuracy of 91% was obtained for logistic regression model.

VI. Application Design

The following diagram describes the process flow. First, we collect the flight information and weather datasets and put them in HDFS. We use Spark-scala to clean the data. The cleaned data is joined. Spark SQL is used for the analytic. Machine learning models are trained to predict the delays using MLlib libraries. Finally, the accuracy for test data is calculated and results are visualized.



The following is application UI which allows the user to enter flight and check if flight is being delayed.
Weather forecast data is collected using NOAA API for the given zip code and date. The data is used as features by the trained model to get the prediction of flight status.



VII. Actuation or Remediation

1. In response to the analytic results, the user can see the monthly distribution of flight delays while planning a trip and keep necessary time buffer for his next task in the itinerary. The user can also choose the airport wisely while booking the flight taking into consideration the delay expected.
2. Before the day of air travel, one can check the prediction 4-5 days ago to check the delay status and take necessary action for his tasks after the travel.

VIII. EXPERMIENTS

The first step was data cleaning. The main goal was to identify the relationship between flight delays and weather and flight information for few airports. The datasets used for the analytic were stored in HDFS. The analysis was performed on NYU HPC cluster. Apache Spark was used as big data processing framework available on HPC cluster. Spark was used for data cleaning. First, we developed the cleaning code and used a subset of data to test the code. Once, the code was working properly, data was ingested in larger batches. Many irregularities were found in the data at this stage when larger dataset was used and the code failed in certain cases. Weather data had many fields missing. Dropping the columns lead to very small size data which was not very useful. So, for wind max and min temperatures average values for that airport was assigned. For missing snow and precipitation, 0 was assigned. For airport, there were missing columns in certain records which were dropped. Some origin and destination airport codes were wrong and such records were dropped as well. Dates were invalid in few records. The cleaning of data required multiple iterations.

Weather and flight datasets were joined on date and airport code to get the flight information and weather for a particular date and airport code. Spark ecosystem tool Spark-SQL was used for analyzing the total delays for every airport, month-wise delays for airports and flight distance wise delays. To predict whether a flight will be delayed Spark-MLlib models were used. Support vector Machines, logistic regression, decision tree and random forest models were used to train the model to use to features and classify if flight will be delayed. To make the class distribution uniform among the records, Synthetic Minority Over-sampling Technique [13] was used. SBT was used to compile scala files and create a jar to run on the spark cluster.

A web application was created using flask which allows the user to input the flight information – airport code, zip code and date till 5 days in future. Weather data forecasted for the date selected is fetched using the NOAA SDK API. This data is used as input by the trained machine learning model to predict if the flight will be delayed.

After doing the delay analysis by airports it was observed the percentage of flights delayed were almost similar for all five airports which lead to conclusion that many factors like propagation delay, arrival delay, airport operations affect the delay along with the weather. Different features were used to see the effect on the delays. Using only weather features did not give significant results. An accuracy of 78% was achieved on test data. Adding flight information like distance and arrival delay the accuracy increased to about 90%. Addition of more features like the delay at the airport, propagation delay and chained flight delay analysis can be done to improve the predictions.

IX. Conclusion

It is concluded that the number of flight delays are positively correlated with the weather conditions. The reason is weather conditions affects the visibility which can lead to accidents and hence it is safer to fly when the visibility is proper.

There are several other factors like arrival delay which is directly related to departure delay and propagation delay caused to aircraft loading, unloading, cleaning and safety inspection which might lead to departure delay. Thus, to capture the complete picture more factors need to be considered. Also, for connected flights, the delay gets propagated for arrival and departure. A chained model which takes into account both the arrival and departure delay can be developed for accurate prediction of delay at the destination.

X. Future Work

In order to get more understanding of data spatial analysis can be done to see the data distribution. Adding more flight related information can help in improving the prediction analysis. As using many features can confuse the classifier only features affecting the analytic most need to be chosen. Random forest algorithm can be used to select features according to the importance and then do the training. A chained delay prediction model can be developed for connected flights. A regression model can be used to predict the delay time.

Acknowledgment

We are thankful to the NYU HPC support for quickly responding to all our questions related to Dumbo and Tableau. Thanks to Cloudera for providing us with access to Hadoop cluster. Thanks Professor Suzanne McIntosh for constantly guiding and encouraging us. We would also like to thank Sree Lakshmi Addepalli for helping us with the blockers and providing references for the project.

References

1. U.S. Department of Transportation. Research and Innovative Technology Administration. Bureau of Transportation Statistics Scheduled Intercity Transportation: Rural Service Areas in the United States. June 2005. Washington, DC: 2005.
2. Chang-Geun. Oh. “Application of Big Data Systems to Aviation and Aerospace Fields; Pertinent Human Factors Considerations,”, International Symposium on Aviation Psychology 2017, May 2017.
3. Alice Sternberg. Jorge Soares. Diego Carvalho. and Eduardo Ogasawara. A Review on Flight Delay Prediction, arXiv:1703.06118 http://arxiv.org/abs/1703.06118, March 2017.
4. Navoneel Chakrabarty. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines, https://arxiv.org/abs/1903.06740, March 2019.
5. Jun Chen. Meng Li. Chained Predictions of Flight Delay Using Machine Learning, AIAA Science and Technology Forum and Exposition 2019, CA, January 2019.
6. Miguel A. Martínez-Prieto. Anibal Bregon. Ivan Garcia. David Scarlatti. Integrating flight-related information into a (Big) data lake, 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), August 2017.
7. National Centers for Environmental Information. National Oceanic and Atmospheric Administration. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00013874/detail
8. Yao Wang. Deepak Kulkarni. Modeling weather impact on ground delay problems, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20110018181.pdf, October 2011.
9. Priyanka Chouksey. Abhishek Singh Chauhan. Weather Data Analytics using MapReduce and Spark, https://ijarcce.com/upload/2017/february-17/IJARCCE%2010.pdf, February 2017.
10. Veershetty Dagade. Mahesh Lagali. Supriya Avadhani. Priya Kalekar. Big Data Weather Analytics Using Hadoop, https://pdfs.semanticscholar.org/f2e4/918444be9b30f29132e93ce02d29ccf26eda.pdf, April 2015.
11. Sun Choi. Young Jin Kim. Simon Briceno. Dimitri Mavris. Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms, https://ieeexplore.ieee.org/document/7777956, September 2016.
12. Ning Yang. Lewis Westfall. Ms. Preeti Dalvi. A weather prediction model with Big Data, http://csis.pace.edu/~ctappert/srd2018/2018PDF/c7.pdf, May 2018.
13. Chawla, N.V. Bowyer, K.W. Hall, L.O. Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 2002, 16, 321–357.