Airline Delay and Cancellation Predictions using Machine Learning

Mukesh Ranjan Sahay
Department of Software Engineering
San Jose State University
San Jose, United States of America
mukesh.sahay88@gmail.com

Muthu Kumar Sukumaran
Department of Software Engineering
San Jose State University
San Jose, United States of America
muthu220515@gmail.com

Sudha Amarnath
Department of Software Engineering
San Jose State University
San Jose, United States of America
sudha04.a@gmail.com

Thirumalai Nambi Doss Palani
Department of Software Engineering
San Jose State University
San Jose, United States of America
ptndoss@gmail.com

***Abstract* — One of the biggest downsides of traveling is the delays and cancellation of the flights which results in the deteriorating customer satisfaction. These delays are associated with many factors including the weather conditions, connecting flights or technical issues, and these prove costly, both quantitative and qualitative.[1] Also, the growing aviation industry has resulted in the air-traffic congestion which increases the chances of the flights getting delayed. Along with the economic impacts, the flight delays are also associated with the harmful effects on the environment. With all these things going around, it is becoming very challenging to effectively manage the air-traffic and to meet the traveler's expectations.**

Therefore, there is an increasing need to anticipate, predict and mitigate the flight delays which can help the airports and the airline companies to improve their performance. Also, these predictions can lead to some customer-oriented measures that can remove or reduce the effects of the delays on their customers. The main objective of this project is to predict the delays of the flights. For this purpose, we will be applying various machine learning algorithms like the decision tree, logistic regression and neural network-based classifiers to predict if a given flight will be delayed or if it will run on time. This implementation of such a predictive model is feasible by applying various Data Mining techniques. There are a lot of variables and data features but finding the variables those have the highest impact on the flight delays is very crucial and for this we will applying the principal component analysis technique. This also improves the performance of the prediction model. Also, we need to find out the patterns and scenarios in which the delays are propagated in the networks of the airports.[2] This is not feasible using the traditional methods and to achieve this we are proposing a prediction model which combines the multi-label random forest classification and approximated delay propagation model.

In this project, we will be applying the Knowledge Discovery Database techniques - data collection, data sampling, data partitioning (training and testing), data pre-processing and data transformation (data normalization and variable selection). Finally, the machine learning models will be applied on these datasets to do the prediction and the evaluation of the model performance using various metrics needs to be done.

For the chained delay prediction, the departure delay and the late arriving flight delay are the most important feature and the expected results of machine learning prediction should follow the same. Also, we are expecting that the variables related to flight departure, weather, flight characteristics and the chained propagation will be the most critical factors for predicting the flight delays. Once we have these data, we will be performing the comparison of the results from various algorithm to find their efficiency.

Keywords— Data Mining; Machine Learning; Predictive Analysis; Flight Delays; Chained Delays; Delay Prediction.

I. INTRODUCTION

Traveling is enjoyable for many of us as it brings joy in meeting new people and visiting adventure places. At times traveling can be extremely stressful because of unprecedented events. Flight delays are one of them. Unpredicted flight delays can ruin the extraordinary vacation memories. The airline delay makes the passengers lose their trust from the famous and internationally recognized airlines. An advanced and automated prediction system with a great accuracy must be created that can predict the likely airline delay. This system can save passengers from the hassle and can also help the airlines to run their business smoothly.

Apart from the delays resulting due to the weather, security and the limited airspace capacity, one-third of the airline delays were caused by a late-arriving aircraft and thus making it depart late for its next schedule, which is known as delay propagation [3, 19]. Generally, the airlines run their aircraft on a scheduled itinerary daily which requires the transit from a network of airports, a late-arriving aircraft early in the day can significantly impact the upcoming flights [2]. For example, if an aircraft is delayed by one hour in a departure from the first airport, it will almost certainly be late in arriving at its next airport; the late arrival may also result in a late subsequent departure of that aircraft, which will lead to a sequence of late-arriving aircraft delays [3].

Here in this paper, we aim is to visualize and predict flight delays with various machine learning technologies and statistical algorithms. We will be applying various machine learning algorithms like the decision tree, logistic regression, and neural network-based classifiers to predict if a given flight will be delayed or if it will run on time. This implementation of such a predictive model is feasible by applying various Data Mining techniques. There are a lot of variables and data features but finding the variables that have the highest impact on the flight delays is very crucial and for this, we will apply the principal component analysis technique. This also improves the performance of the prediction model. Also, we need to find out the patterns and

scenarios in which the delays are propagated in the networks of the airports.[2] This is not feasible using the traditional methods and to achieve this we are proposing a prediction model that combines the multi-label random forest classification and approximated delay propagation model.

Below are the categories we incorporated to handle the data from multiple sources, visualize and apply machine learning algorithms to predict the best possible outcome. Reasons for the delays could be related to – weather, security, air system or late aircraft arrivals.
Following steps and procedures are flowed through the course of this paper – Data gathering and cleaning, delay comparison for the airlines, delay in takeoff and landing, the relationship between the source and destination airports, frequency of delay for the airlines and predicting the delay for one airline compared to others.

The outline of the rest of the paper is as follows: Section 2 deals with the literature review for the topic. Section 3 describes the related works done in this area which is the State-of-the-Art review on this topic. Section 4 lists downs the data sources used for this paper and explains the data in detail. Also, it explains the data mining techniques to be applied and provides the list of selected features for analysis, and how we selected those features using the Principal Component Analysis. Section 5 explains the machine learning algorithms used for this research and, describes the implementation of the model. Section 6 describes the results of various algorithms and a comparison of these results. Also, this section will be used to present the performance metrics of various algorithms Finally, section 7 presents and proposes the best algorithm for the flight delay prediction and the related future work that can be done in this area.

II. LITERATURE REVIEW

Chained predictions of flight delay using machine learning [1]
Methods are needed to analyze the manner how delay propagates in the airport networks. Traditional methods are inadequate to the task. This paper presented a new machine learning based air traffic delay prediction model that combined multi-label random forest classification and approximated delay propagation model.

Prediction of weather-induced airline delays based on machine learning algorithms [5]
Flight schedule and weather forecast were gathered and fed into the model to determine the flight delay.

A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines [6]
In this paper, a two-stage predictive model was developed employing supervised machine learning algorithms for the prediction of flight on-time performance. The first stage of the model performs binary classification to predict the occurrence of flight delays and the second stage does regression to predict the value of the delay in minutes.

Identifying flight delay patterns using diverse subgroup discovery [7]
In this paper author has applied diverse SD to historic flight data and mine subgroups of flights that, on average, have a large delay. This approach gives subgroups that can be easily understood by experts, even though non-trivial relations between multiple variables can be discovered. Using diverse SD gives less redundant results than standard top-k SD and demonstrate that even in situations where inferring an accurate predictive model is infeasible, local deviations can be effectively captured and described by local patterns, potentially providing valuable insights for, e.g., airline scheduling problems.

A deep learning approach to flight delay prediction [8]
This paper investigates the effectiveness of the deep learning models in the air traffic delay prediction tasks. In this study, four different ways of building deep RNN architecture are also discussed. Finally, the accuracy of the proposed prediction model was measured, analyzed and compared with previous prediction methods. It shows best accuracy compared with all other methods.

III. DATASETS AND FEATURES

A. Source of the Data

The data for this research is gathered from the Bureau of Transportation and Statistics (BTS) which has three different data points - Airline On-Time Dataset, Airport Dataset and Flights. The airline on-time data are reported to the Department of Transportation, BTS in 2017 by the 12 largest airlines of the United States. This data covers non-stop scheduled flights between airports within the United States. The other two datasets provide information about the on-time performance of the flights operated by major carriers including the summary of the information on the number of on-time, delayed, canceled and diverted flights.

B. Data Mining Techniques

Data mining is the technique of validating the data, looking for the hidden and potentially useful/meaningful patterns in large datasets. Data Mining is about discovering the previously unknown relationships and information from the data. It is a multi-disciplinary skill that uses machine learning, statistics, artificial intelligence and database technologies. The insights deduced using the Data Mining can be used for sales, marketing, anomaly detection, and scholastic discoveries. Data mining is also known as the knowledge discovery/extraction process, data and pattern analysis.

Various data mining techniques are – Classification, Clustering, Regression, Association rules, Outlier detection, Sequential pattern and Prediction.

Classification is used to fetch important information about data, and its metadata. It is used to classify the data in different categories. Clustering is a data mining technique for identifying similar data, by helping to understand the similarities and uniqueness of the data. Regression is the process of identifying/analyzing the relationship between

variables i.e. the likelihood of occurrence of one feature as compared to other features of the datasets. The association rules help to find the association between two or more data in the datasets and help to discover the unidentified patterns in the dataset. Outlier detection is used to identify the data in the dataset which does not match the expected pattern/behavior. Sequential patterns are used to discover the similar trends in the data for a given period. The prediction uses a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. to analyze the existing data and then make a further prediction.

C. Training Data and Feature Selection

The datasets mentioned above were amalgamated in order to perform analysis on them. Later, the same dataset was split into two – the training dataset (80%) and the test dataset (20%).

All the features in the airline dataset are highly significant, still, some features are eliminated from the dataset due to various reasons:

- Year is a categorical data having almost no variability as the dataset contains flight information of only specific years. Therefore, this feature can be dropped from the dataset.
- Quarter is also not that important feature and can be dropped, as the same information can be determined using the month feature in the dataset.

Also, other features, like F8 and F10 seems to be repeated since the Origin and Destination Airports are already determined using the F7 and F9 respectively from the dataset. Apart from the dataset under consideration, there are occurrences where more than one airport may have the same code and so in order to preserve the consistency of the model, F8 and F10 are not dropped. Below is the heatmap presenting the correlation matrix for displaying the Feature-to-Feature and Feature-to-Label correlations where all the features are continuous.

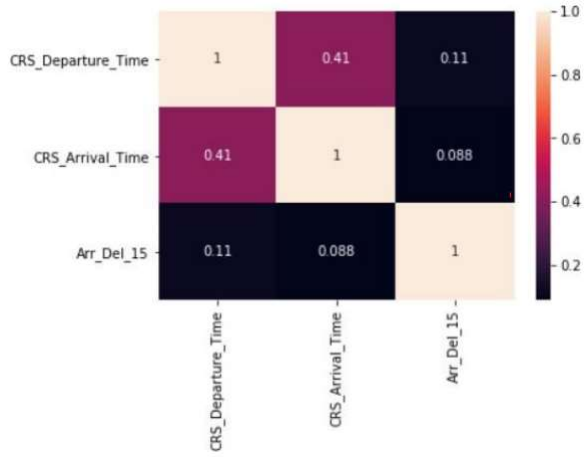

Figure 1. Heat Map Showing Correlation

D. Principal Component Analysis

The principal component analysis is a technique to analyze data under observation which is described by several correlated variables. The purpose of PCA is to extract the dominant and important features from the data, which can be presented and used for further analysis. The accuracy of the PCA model is determined using the cross-validation techniques like the jackknife and bootstrap, which depends on the scaling of the dataset. In the beginning, it looks for outliers and strong groupings in the data plots, indicating that the data enrichment should be performed.

E. Data Preprocessing

Before applying the machine learning algorithms for training the model on the datasets, we need to apply some data preprocessing and cleansing techniques so that accurate results are obtained.

Following are the high-level steps that are followed:

- Handling the missing values in the dataset
- Adding labels to the categorical features using label encoding.
- One-Hot encoding for splitting the different categorical features into categories and assigning the binary values.
- Data imbalance removal and balancing the data

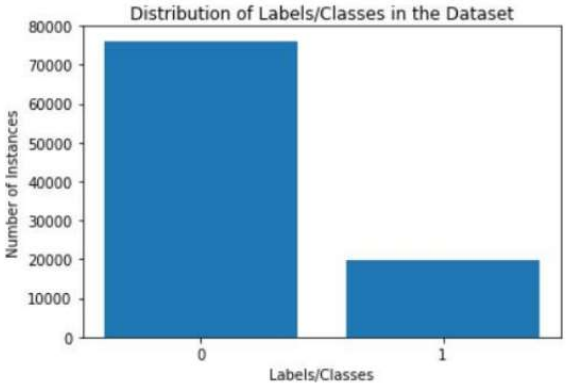

Figure 2. Data Imbalance

IV. Machine Learning Algorithms and Methods

A. One-Hot Encoding

One-Hot encoding is one of the processes of converting categorical variables into a form that could be injected into the machine learning algorithms to do prediction with increased accuracy. In the case of categorical data, there is no ordinal relationship, and in such cases, the integer encoding is not enough. Indeed, using integer encoding and letting the model consider the default ordering between the categories may result in degraded performance or an unexpected result. In these cases, the one-hot encoding should be used on the integer representation, during which the integer encoded variables are replaced with a new binary variable with a unique integer value.

B. Linear Regression

Linear regression is used to analyze the relationship between two different variables by applying a linear equation to input

data. One of the variables is used as an explanatory variable, and the other is the dependent variable. For example, the relationship between the airline delay and time of the departure. It is used to determine the statistical relationship, not the deterministic relationship.

In order to fit a linear regression model to a dataset, we should first determine if there is a relationship between the features. This does not mean we are looking for a variable that leads to other, instead, we should look for some significant association between the two variables. There are few tools that help to determine the strength of the relationship of the features like a scatterplot. If there is no association between the two features, then fitting a linear regression model to that data using these features will not provide useful information. The correlation coefficient is a valuable measure of association between two variables, which is any value from -1 to 1 which indicates the strength of the association for the two variables.

C. Neural Network

Neural networks are nothing but a set of algorithms, which is modeled loosely based on the human brain, and are designed to identify the patterns. They are used to interpret the sensory data using a type of machine perception, labeling or clustering raw data. The recognized patterns are numerical, which are contained in the vectors, into which all the real-world data like images, sound, text or time series must be converted. Neural networks help to do clustering and classification of the data. It can be considered as a clustering and classification layer on top of the data under analysis. They help in grouping the unlabeled data depending on the similarities with the sample input and then classify those data to train the model using it. Neural networks can also help extract features that are injected into other machine learning algorithms. So, it can be considered as the components of bigger machine-learning systems that involves the algorithms for reinforcement learning, classification, and regression.

D. Random Forest Classification

The Random Forest (RF) classifier is an ensemble method based on multiple decision trees [12]. By combining the Bootstrap aggregating [13]and random space method [14], RF overcomes the drawbacks of individual decision tree. RF is widely used in industry because it can classify high dimensional data in short time with good performance and it has low sensitivity to outliers in the training data [13]. Moreover, RF was chosen as the core for our prediction modules for two reasons. First, RF is tested to have superior performance than other classification models [10,11]. Second, RF can output the importance of the features in its learning process, which is the key for our feature selection process.

E. Approximated Delay Propagation Model

Delay propagation of the flights and in the air traffic network has a very random nature due to the multiple deciding factors in the system and is like the process of epidemic spreading. So, an air traffic network can be modeled as a directed graph with the nodes and edges by considering the various factors of the air transportation system like airports, airspaces, flight routes, and others. Two different models can be constructed using the airport-based and flight-based concepts.

A mixed approach for chained delay prediction can be modeled as shown in figure 3. In this system, the arrival delay is predicted and injected to the system for predicting the departure delay. The predicted departure delay is in turn used to improve the arrival delay prediction on the selected features from the datasets.

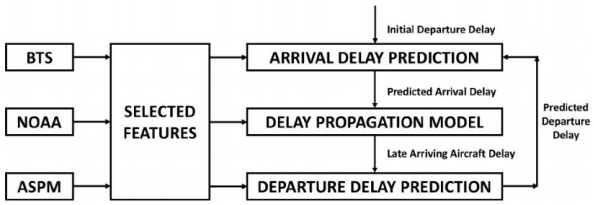

Figure 3. Mixed approach for chained delay prediction

V. EXPERIMENT RESULTS

The flight data available from the data set can be merged to airport dataset and can be visualized in map using BaseMap libraries.

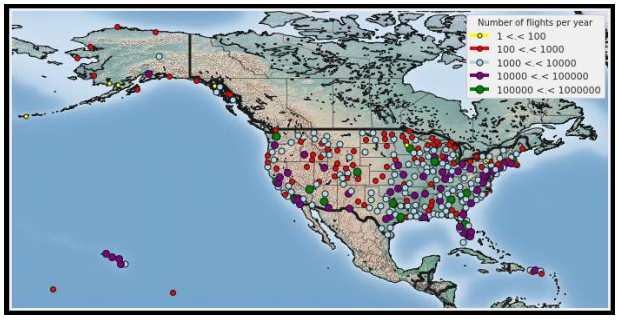

Figure 4. Number of flights per year at different airports

Data cleaning and depth analysis of the data indicates that the ARRIVAL_DELAY is 97% empty. Since the percentage of data is huge for mockup or imputing, we decided to drop the column as considering this will cause heavy load to unknown feature.

Following figures shows the result of the delay per airline carrier and their overall percentage.

Figure 5. Legends for the Airline Carrier

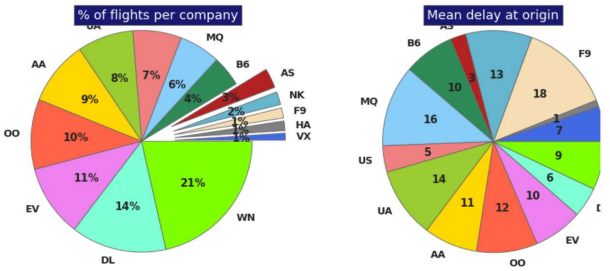

Figure 6. Overall share and delay at origin

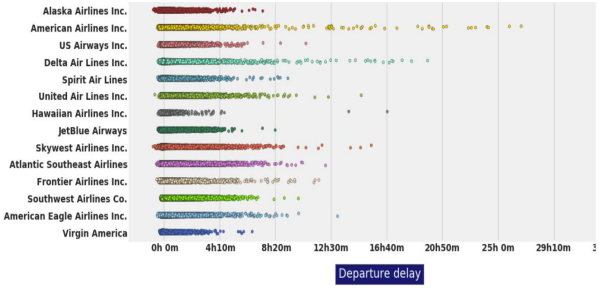

Figure 7. Airline delay at departure

Next, we analyzed the duration of the delay for different airlines and its distribution. The result is as below.

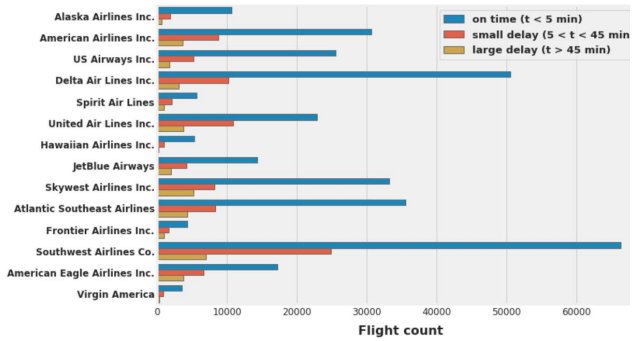

Figure 8. Delay duration

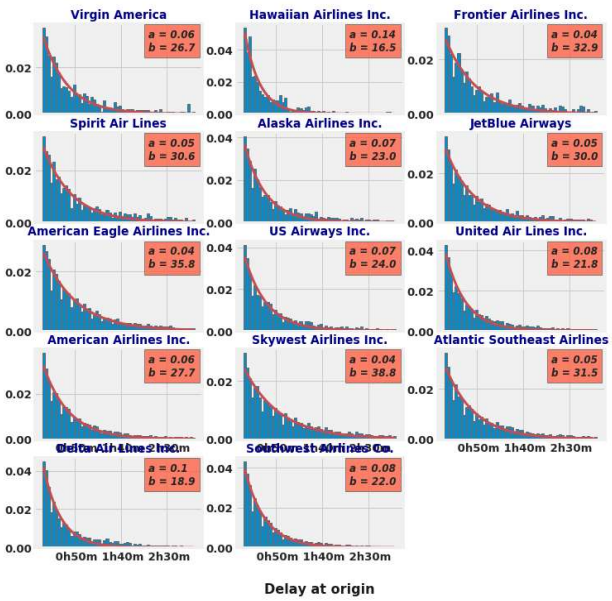

Figure 9. Delay distribution among airlines

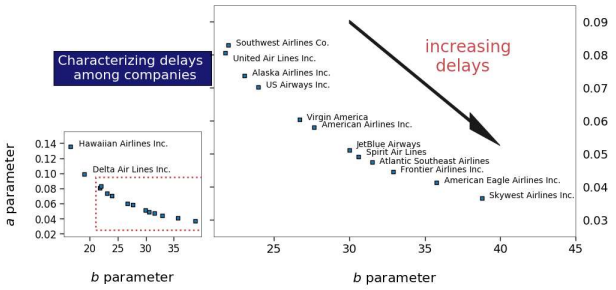

Figure 10. Delay Comparison

Comparison of the delay based on the takeoff or landing.

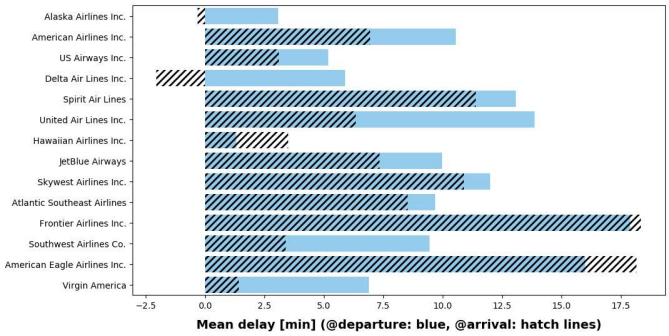

Figure 11. Takeoff vs Landing

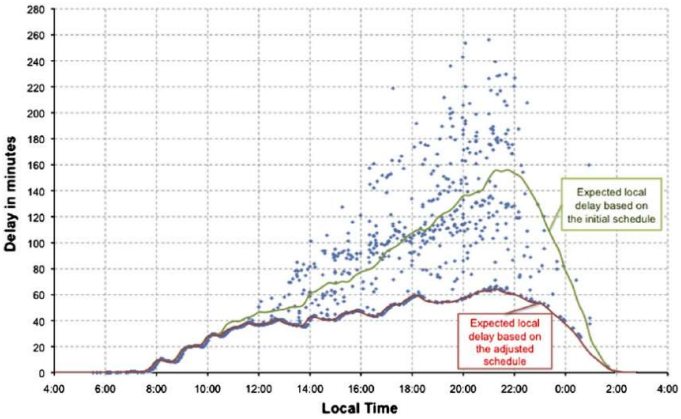

Figure 12. Arrival delay compared to the expected time

After these, we performed the analysis based on the airports data and plotted the number of airports visited by each airline.

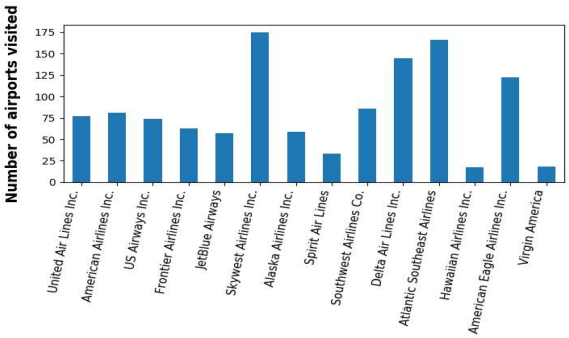

Figure 13. Number of airports visited by airline

The same is plotted on the map for better understanding for few of the airline carriers.

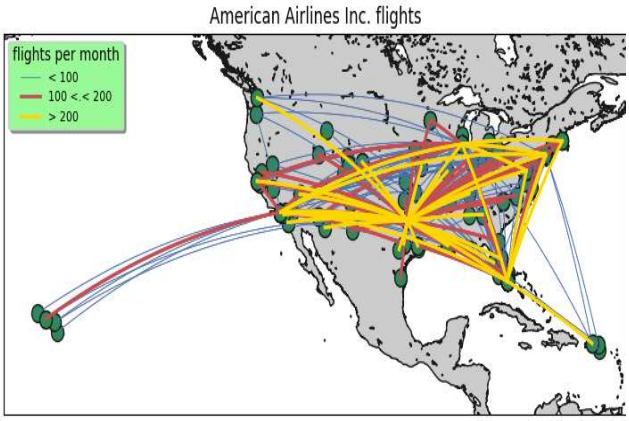

Figure 14. American Airlines

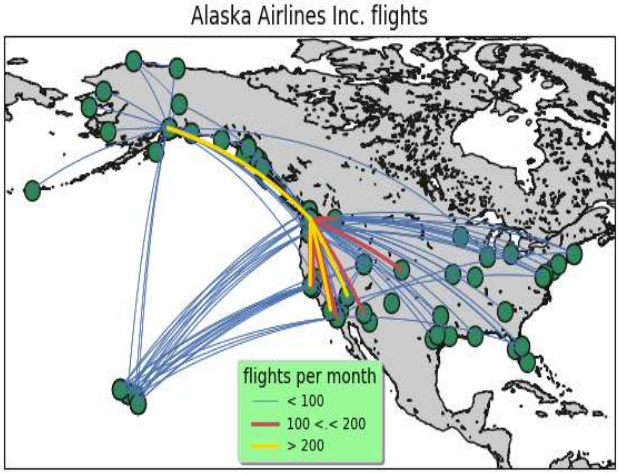

Figure 15. Alaska Airlines

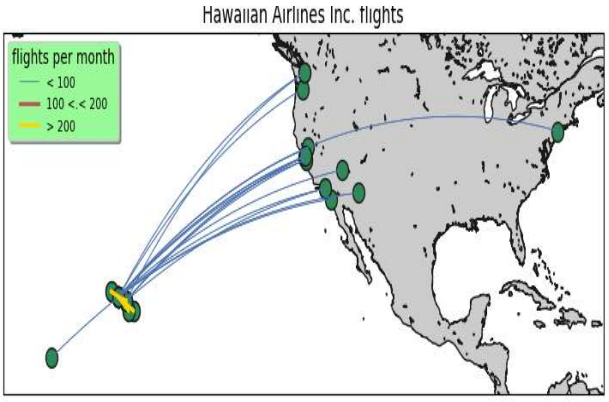

Figure 16. Hawaiian Airlines

Analysis of the impact of the originating airport on the flight delay.

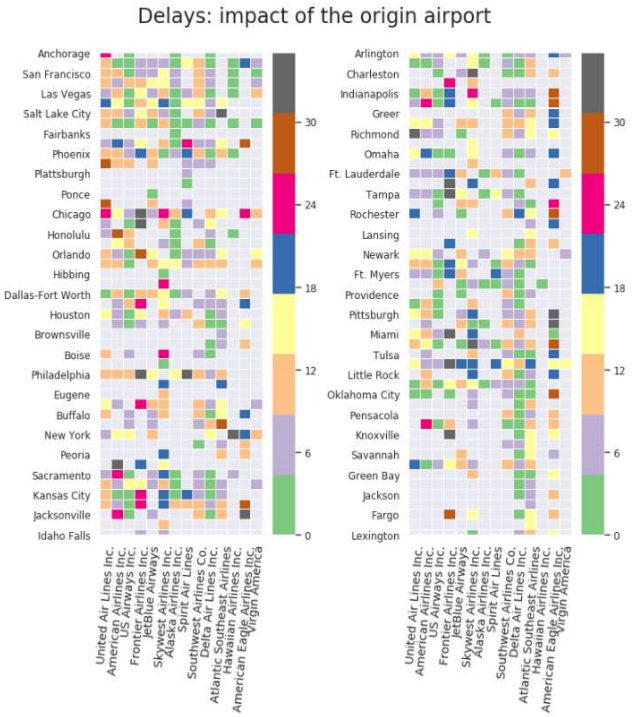

Figure 17. Impact of origin airport on delay

We did an analysis of the delay depending on the departure date and time. The results are:

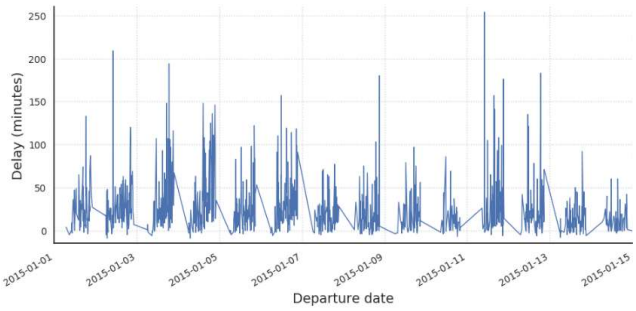

Figure 18. Delay vs Departure Date

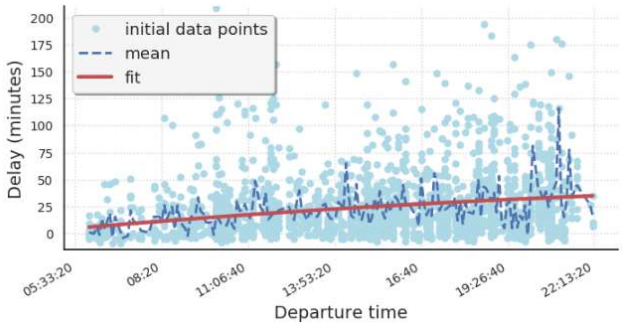

Figure 19. Delay vs Departure Time

We applied various machine learning algorithms for predicting the flight delay based on the datasets under analysis. The results are shown below:

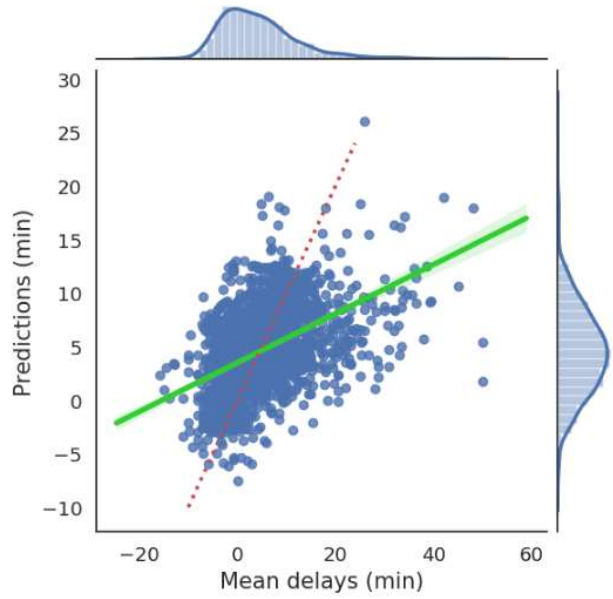

Figure 20. One-Hot Encoding Prediction

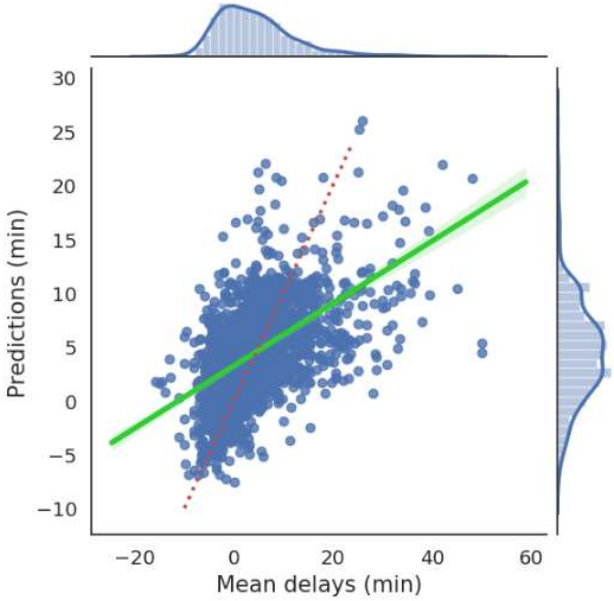

Figure 21. Prediction using Linear Regression

VI. Conclusion and Future Work

During this research, we applied various machine learning algorithms on the airline, airport and the flights datasets in order to determine the relationship of various factors with the flight delay and cancellation. Also, we applied various models including the one-hot encoding, linear regression for predicting the airline delay. Based on the results, we can conclude that the flight delays can be introduced due to various factors like weather, air traffic, airport location and activities, but the major impact is of the chained delay arising due to the late arriving flights which is propagated to the airport network. The flight delay is also impacted to a greater extent by the originating airport and the airline carrier.

Future works include analyzing the weather and environmental datasets and make use of those datasets to predict the flight delays. Also, datasets related to the airport staff and functionalities can be introduced into the prediction model. Apart from this, various other machine learning and deep learning algorithms are available to be applied for prediction and a comparison can be done to determine the best model depending on the accuracy of the prediction.

Acknowledgment

We thank our professor Chandrasekar Vuppalapati from San Jose State University who provided us this opportunity to work on this research paper and project and guided us on the same. We would also like to show our gratitude to our friends for sharing their pearls of wisdom with us during this research and thank them for reviewing this paper.

References

1. Feiteira I. (2018). Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport. Procedia Computer Science,138,638-645.

2. Chen, Jun & Li, Meng. (2019). Chained Predictions of Flight Delay Using Machine Learning. 10.2514/6.2019-1661.

3. N. Pyrgiotis, K. M. Malone, and A. Odoni, “Modelling delay propagation within an airport network,” Transportation Research Part C: Emerging Technologies, vol. 27, pp. 60–75, 2013.

4. K. Gopalakrishnan and H. Balakrishnan, “A comparative analysis of models for predicting delays in air traffic networks,” in USA/Europe Air Traffic Management Seminar , 2017

5. Sun Choi (Aerosp. Syst. Design Lab., Georgia Inst. of Technol., Atlanta, GA, United States); Young Jin Kim; Briceno, S.; Mavris, D. Source: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). Proceedings, p 6 pp., 2016

6. Thiagarajan, B. (Sri Venkateswara Coll. of Eng., Chennai, India); Srinivasan, L.; Sharma, A.V.; Sreekanthan, D.; Vijayaraghavan, V. Source: 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). Proceedings, p 6 pp., 2017

7. Proenca, Hugo M. (LIACS, Leiden University, Leiden, Netherlands); Klijn, Ruben; Bäck, Thomas; Van Leeuwen, Matthijs Source: Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018, p 60-67, July 2, 2018, Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018

8. Kim, Young Jin (Aerospace Systems Design Laboratory, Georgia Institute of Technology, Atlanta; GA; 30332-0150, United States); Choi, Sun; Briceno, Simon; Mavris, Dimitri Source: AIAA/IEEE Digital Avionics Systems Conference - Proceedings, v 2016-December, December 7, 2016, 35th DASC Digital Avionics Systems Conference 2016, DASC 2016 – Proceedings

9. Chakrabarty, Navoneel, et al. ”Flight Arrival Delay Prediction Using Gradient Boosting Classifier.” Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 651-659.

10. Suvojit Manna, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, Subhas Burman ”A statistical approach to predict flight delay using gradient boosted decision tree”, International Conference on Computational Intelligence in Data Science(ICCIDS), 2017

11. Juan Jose Robollo and Hamsa Balakrishnan ”Characterization and Prediction of Air Traffic Delays”

12. Yi Ding ”Predicting flight delay based on multiple linear regression”, IOP Conference Series: Earth and Environmental Science.

13. Sruti Oza, Somya Sharma, Hetal Sangoi, Rutuja Raut, V.C. Kotak ”Flight Delay Prediction System Using Weighted Multiple Linear Regression”, International Journal Of Engineering And Computer Science ISSN:2319-7242, Volume 4 Issue 4 April 2015, Page No. 11668-11677

14. Anish M. Kalliguddi and Aera K. Leboulluec ”Predictive Modeling of Aircraft Flight Delay”, Universal Journal of Management 5(10): 485-491, 2017, DOI: 10.13189/ujm.2017.051003

15. Jianmo Ni, Xinyuan Wang, Ziliang Li ”Flight Delay Prediction using Temporal and Geographical Information”,https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a032.pdf

16. Dong, Yanjie, and Xuehua Wang. ”A new over-sampling approach: random-SMOTE for learning from imbalanced data sets.” International Conference on Knowledge Science, Engineering and Management. Springer, Berlin, Heidelberg, 2011.

17. Li, Jia, Hui Li, and Jun-Ling Yu. ”Application of random-SMOTE on imbalanced data mining.” Business Intelligence and Financial Engineering (BIFE), 2011 Fourth International Conference on. IEEE, 2011.

18. Sina Khanmohammadi, Salih Tutun, Yunus Kucuk ”A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport”, doi.org/10.1016/j.procs.2016.09.321

19. Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A. A., and Zou,B., “Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States,NEXTOR,” 2010.