PREDICTING FLIGHT DELAYS USING MACHINE LEARNING

Tushar Patel
UG Scholar, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Nitish Singh
UG Scholar, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Chintala Chereesh
UG Scholar, Computer Engineering Department, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

DR. G. Somasekhar
Associate Professor, CMR Technical Campus, UGC Autonomous, Kandlakoya (V), Medchal Road, Hyderabad-501401, INDIA

Abstract:

Flight delay is a major problem in the aviation sector. During the last two decades, the growth of the aviation sector has caused air traffic congestion, which has caused flight delays. Flight delays result not only in the loss of fortune also negatively impact the environment. Flight delays also cause significant losses for airlines operating commercial flights. Therefore, they do everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In this paper, using machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.

Keywords*—Flight Prediction, Machine Learning, Logistic Regression, U.S. Flight data.*

I Introduction

The growing aviation industry has resulted in air-traffic causing flight delays. Flight delays have huge economic impact on airlines and also have harmful environmental effects. Therefore, it is important to detect the major factors influencing flight delay. The factors influencing the flight delay range from natural factors like weather to factors like day, month, etc. On observing wide variety of flight data, a list of factors responsible for the delay was generated. There are other factors influencing the delay as well, but their scope is limited. This list primarily categorizes delay into two types, namely departure delay and arrival delay. Some of these features influence the flight delay drastically while others have a minor impact. In order to develop an efficient flight delay prediction system, these features along with their degree of impact on the delay must be taken into consideration.

For all airlines, flight delays represent the source of financial and technical difficulties. This article aims both to identify and analyse the factors causing delays and to suggest some possibilities on how to eliminate these delays.

Flight delay is studied vigorously in various research in recent years. The growing demand for air travel has led to an increase in flight delays. According to the Federal Aviation Administration (FAA), the aviation industry loses more than $3 billion in a year due to flight delays [1] and, as per BTS [2], in 2016 there were 860,646 arrival delays. The reasons for the delay of commercial scheduled flights are air traffic congestion, passengers increasing per year, maintenance and safety problems, adverse weather conditions, the late arrival of plane to be used for next flight [3] [4]. In the United States, the FAA believes that a flight is delayed when the scheduled and actual arrival times differs by more than 15 minutes. Since it becomes a serious problem in the United States, analysis and prediction of flight delays are being studied to reduce large costs.

II. LITERATURE SURVEY

Much research has been done on studying flight delays. The prediction, analysis and cause of flight delays have been a major problem for air traffic control, decision-making by airlines and ground delay response programs. Studies are conducted on the delay propagation of the sequence. Also, studying the predictive model of arrival delay and departure delay with meteorological features is encouraged. In the past, researchers have tried to predict flight delays with Machine Learning. Chakrabarty et al. [5] used supervised automatic learning algorithms (random forest, Gradient Boosting Classifier, Support Vector Machine and the k-nearest neighbour algorithm) to predict delays in the arrival of operated flights including the five busiest US airports. The maximum precision achieved was 79.7% with gradient booster as a classifier with a limited data set. Choi et al. [6] applied machine learning algorithms like decision tree, random forest, AdaBoost and kNearest Neighbours to predict delays on individual flights. Flight schedule data and weather forecasts have been incorporated into the model. Sampling techniques were used to balance the data and it was observed that the accuracy of the classifier trained without sampling was more that of the trained classifier with sampling techniques. Cao et al. [7] used a Bayesian Network model to analyse the turnaround time of a flight and delay prediction.

Juan José Rebollo and Hamsa Balakrishnan [8] used a hundred pairs of origin and destination to summarise the result of various regression and classification models. The find outs reveal that among all the methods used, random forest has the highest performance. However, predictability may additionally range because of factors such as the number of origindestination pairs and the forecast horizon. Sruti Oza, Somya Sharma [9] used multiple linear regression to predict weatherinduced flight delays in flight-data, as well as climatic factors and probabilities due to weather delays. The forecasts were based on some key attributes, such as carrier, departure time, arrival time, origin and destination. Anish M. Kalliguddi and Aera K. Leboulluec [10] predicted both departure and arrival delays using

regression models such as Decision Tree Regressor, Multiple Linear Regression and Random Forest Regressor in flight-data. It has been observed that the longer forecast horizon is useful for increasing the accuracy with a minimum forecast error for random forests. Etani J Big Data [11] A supervised model of on-schedule arrival fight is used using weather data and flight data. The relationship between flight data and pressure patterns of Peach Aviation is found. On-Schedule arrival flight is predicted with 77% accuracy using Random Forest as a Classifier.

III. PROPOSED METHODOLOGY

To predict flight delays to train models, we have collected data accumulated by the Bureau of Transportation, U.S. Statistics of all the domestic flights taken in 2015 was used. The US Bureau of Transport Statistics provides statistics of arrival and departure that includes actual departure time, scheduled departure time, scheduled elapsed time, wheels-off time, departure delay and taxi-out time per airport. Cancellation and Rerouting by the airport and the airline with the date and time and flight labelling along with airline airborne time are also provided. The data set consists of 25 columns and 59986 rows. The data shows some of the fields of the original dataset. There were many lines with missing and null values. The data must be pre-processed for later use. The methodology here uses the supervised learning technique to gather the advantages of having the schedule and real arrival time. Initially, some specific monitoring algorithms with a light computation cost were considered candidates and therefore the best candidate was perfected for the final model. We develop a system that predicts for a delay in flight departure based on certain parameters. We train our model for forecasting using various attributes of a particular flight, such as arrival performances, flight summaries, origin/destination, etc.

Before applying algorithms to our data set, we need to perform a basic pre-processing. Data preprocessing is performed to convert data into a format suitable for our analysis and also to improve data quality since real-world data is incomplete, noisy and inconsistent. We have acquired a data set from the Bureau of Transportation for 2015. The data set consists of 25 columns and 59986 rows. There were many rows with missing and null values. The data set was cleaned up using the pandas' dropna() function to remove rows and columns from the data set consisting of null values. After preprocessing, the rows were reduced to 54486. Fig. 2 shows the number of records which were null for specific attributes, e.g. there were 1413 records which have null value for attribute.

This project architecture shows the procedure followed for flight delay using machine learning, starting from input to final prediction.

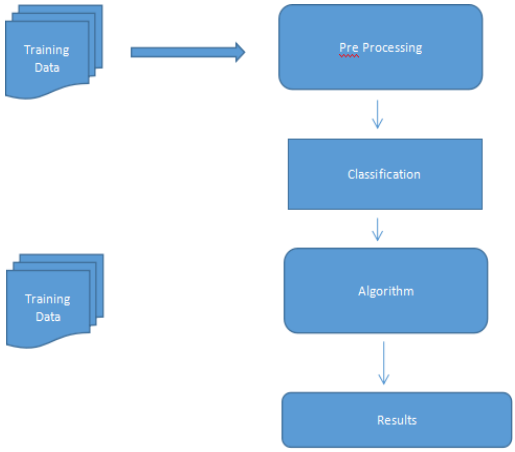

Fig 1: Project Architecture

IV RESULT ANALYSIS

After preprocessing and feature extraction of our dataset, 80% of the dataset was selected for training and 20% of the dataset was selected for testing. For error calculation, we are using scikit-learn metrics.

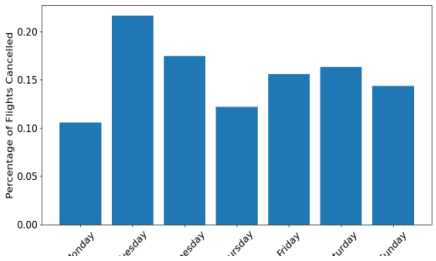

Fig 2: Flights canceled on the weekdays

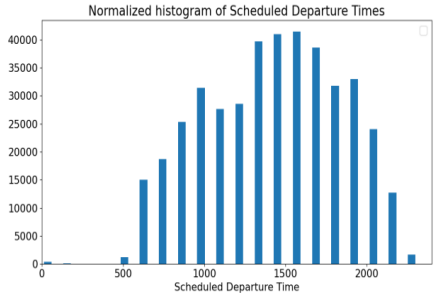

Fig 3: Departures of the scheduled of the flights showed in the histogram

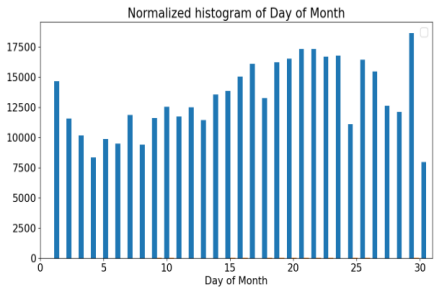

Fig 4 : Heavy traffic categorized based on the month statistics

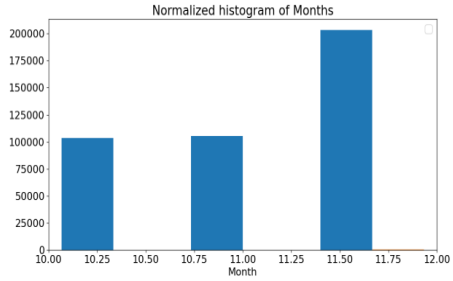

Fig 5 : Peak flight departures from an airport

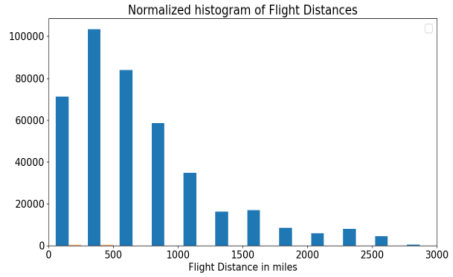

Fig 6 : Flight distances travelled based on the miles

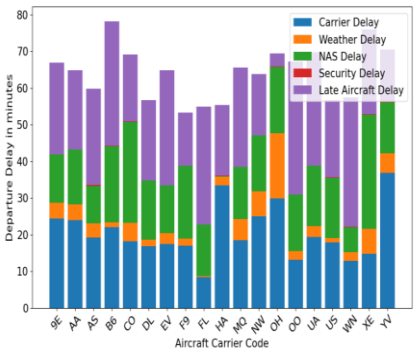

Fig 7 :Departure delay in minutes from different aircraft's

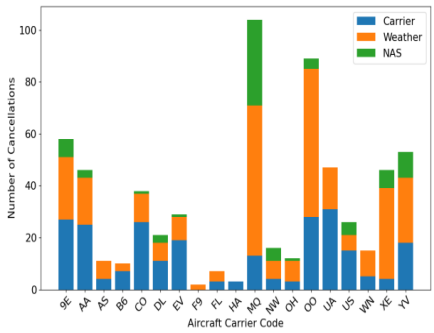

Fig 8 :Number of cancellations based on the different factors

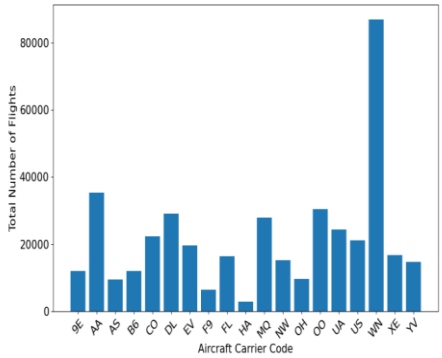

Fig 9 : Total number of flights operated from the source location

```
>>>
>>> X = df[['Month', 'DayofMonth', 'DayOfWeek', 'CRSDepTime', 'Origin', 'Dest', 'Distance', 'Carrier mean distance',
...         'Origin Delay', 'Origin TaxiOut']]
>>> y = df['Cancelled']
>>> from sklearn.model_selection import train_test_split
>>> from sklearn.preprocessing import MinMaxScaler
>>> from sklearn.linear_model import LogisticRegression
>>> from sklearn.model_selection import GridSearchCV
>>> from sklearn.neural_network import MLPClassifier
>>> from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
>>>
>>>
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 0)
>>> scaler = MinMaxScaler()
>>> X_train_scaled = scaler.fit_transform(X_train)
>>> X_test_scaled = scaler.transform(X_test)
>>>
>>> nnclf = MLPClassifier(hidden_layer_sizes = [5,5], solver='adam', alpha=0.0003, activation='relu',
...                       max_iter = 100, random_state = 47).fit(X_train_scaled, y_train)
>>> y_predicted = nnclf.predict(X_test_scaled)
>>> confusion = confusion_matrix(y_test, y_predicted)
>>> print('Accuracy: {:.3f}'.format(accuracy_score(y_test, y_predicted)))
Accuracy: 0.999
```

Fig 10 :MLP classifier performance on the data set

```
>>> from sklearn.ensemble import RandomForestClassifier
>>> from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
>>>
>>> clf = RandomForestClassifier(n_estimators=50, random_state=47).fit(X_train, y_train)
>>>
>>> # sum(y_test)
>>> # clf.score(X_test, y_test)
>>>
>>> y_predicted = clf.predict(X_test)
>>> confusion = confusion_matrix(y_test, y_predicted)
>>> #confusion
>>> #sum(y_predicted)
>>>
>>> print('Recall: {:.3f}'.format(recall_score(y_test, y_predicted)))
Recall: 0.000
>>> print('Precision: {:.3f}'.format(precision_score(y_test, y_predicted)))
Precision: 0.000
>>> print('Accuracy: {:.3f}'.format(accuracy_score(y_test, y_predicted)))
Accuracy: 0.999
```

Fig 11 : Random Forest classifier performance on the data set

```
>>> from sklearn.svm import SVC
>>> np.set_printoptions(formatter={'float': lambda x: "{0:0.3f}".format(x)})
>>>
>>> svm = SVC(kernel='rbf', C=1000, gamma=6, random_state=47).fit(X_train_scaled, y_train)
>>> y_pred = svm.predict(X_test_scaled)
>>>
>>> print('Recall: {:.3f}'.format(recall_score(y_test, y_pred)))
Recall: 0.022
>>> print('Precision: {:.3f}'.format(precision_score(y_test, y_pred)))
Precision: 0.012
>>> print('Accuracy: {:.3f}'.format(accuracy_score(y_test, y_pred)))
Accuracy: 0.996
```

Fig 12 : SVC Classifier performance on the dataset

V CONCLUSION

The primary goal of this project is to predict airline delays caused by various factors and Error rate on models. Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. To carry out the predictive analysis, which encompasses a range of statistical techniques from supervised machine learning and, data mining, that studies current and historical data to make predictions or just analyze about the future delays, with help of Regression Analysis using regularization technique in Python.

VI REFERENCES

[1] N. G. Rupp, "Further Investigation into the Causes of Flight Delays," in Department of Economics, East Carolina University, 2007.

[2] "Bureau of Transportation Statistics (BTS) Databases and Statistics," [Online]. Available: http://www.transtats.bts.gov.

[3] "Airports Council International, World Airport Traffic Report," 2015,2016.

[4] E. Cinar, F. Aybek, A. Caycar, C. Cetek, "Capacity and delay analysis for airport manoeuvring areas using simulation," Aircraft Engineering and Aerospace Technology, vol. 86, no. No. 1,pp. 43-55, 2013.

[5] Navoneel, et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient

Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.

[6] Y. J. Kim, S. Briceno, D. Mavris, Sun Choi, "Prediction of weatherinduced airline delays based on machine learning algorithms," in 35th Digital Avionics Systems Conference (DASC), 2016.

[7] W.-d. Cao. a. X.-y. Lin, "Flight turnaround time analysis and delay prediction based on Bayesian Network," Computer Engineering and Design, vol. 5, pp. 1770-1772, 2011.

[8] J. J. Robollo, Hamsa, Balakrishnan, "Characterization and Prediction of Air Traffic Delays".

[9] S. Sharma, H. Sangoi, R. Raut, V. C. Kotak, S. Oza, "Flight Delay Prediction System Using Weighted Multiple Linear Regression," International Journal of Engineering and Computer Science, vol. 4, no. 4, pp. 11668 - 11677, April 2015.

[10] A. M. Kalliguddi, Area K., Leboulluec, "Predictive Modelling of Aircraft Flight Delay," Universal Journal of Management, pp. 485 491, 2017.

[11] Noriko, Etani, "Development of a predictive model for on-time arrival fight of airliner by discovering correlation between fight and weather data," 2019.

[12] C. J. Willmott, Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square(RMSE) in assessing average model performance," Climate Research, vol. 30, no. 1, pp. 79 - 82, 2005.