Flight Delay Prediction

Bindu Parvati Jonnala Gadda, Pooja Reddy Gummakonda, Sriram Chowdary Yalavarthi, and Vineesha Challagulla

University of Missouri-Kansas City

Kansas City, MO, USA

Abstract*—*A significant issue in the aviation industry is flight delays. The expansion of the aviation industry during the past two decades has increased air traffic, which has delayed flights. Not only do flight delays cost money, but they also have an adverse effect on the surroundings. Airlines that operate commercial flights suffer huge losses as a result of flight delays. In order to prevent or avoid flight delays and cancellations, they consequently employ every feasible measure. In this research, we forecast whether a specific flight's arrival will be delayed or not using machine learning models, including linear regression and random forest regression.

Keywords*—*flight prediction, machine learning, EDA, data pre-processing/data wrangling, 3-sigma rule, linear regression, random forest regression.

I. INTRODUCTION

In recent years, numerous studies have focused heavily on flight delays. There are more flights being delayed because of the rising demand for air travel. Flight delays cost the aviation sector more than $3 billion annually, according to the Federal Aviation Administration (FAA) [1], and according to BTS [2], there were 860,646 arrival delays in 2016. Commercial planned flight delays can occur for a variety of reasons, such as air traffic congestion, an increase in passengers each year, maintenance and safety difficulties, bad weather conditions, and the delayed arrival of the aircraft to be used for the subsequent trip [3] [4]. When a flight's scheduled arrival time and actual arrival time diverge by more than 15 minutes, the FAA in the US considers the flight to be delayed. Analysis and forecasting of flight delays are being researched to save significant expenditures as they become a severe issue in the United States. The goal of the research is to identify the cause of the flight delay in order to enhance safety, passenger satisfaction, airline efficiency, airport capacity planning, and data-driven decision-making.

II. LITERATURE SURVEY

Flight delays have been the subject of extensive research. Air traffic control, airline decision-making, and ground delay response programs have all experienced significant difficulties with the forecasting, analysis, and causation of aircraft delays. On the sequence's delay propagation, research is being done. Additionally, research into the arrival delay and departure delay forecast models using meteorological variables is acknowledged. In the past, researchers have experimented with using machine learning models to forecast aircraft delays. One hundred pairings of origin and destination were used by Juan José Rebollo and Hamsa Balakrishnan [5] to summarize the findings of several regression and classification models. The results show that random forest has the best performance out of all the approaches used. Predictability, however, may also be due to elements like the quantity of origin-destination pairs and the forecasting horizon. Oza Sruti, Somya Sharma [6] employed multiple linear regression to forecast climatic variables and weather-related aircraft delays in flight data and the chances brought on by weather delays. The predictions were based on a few crucial factors, including airline, departure, and arrival times, as well as origin and destination.

Anish M. Kalliguddi and Aera K. Leboulluec [7] used regression models including decision tree regression, multiple linear regression, and random forest regression in flight-data to forecast both departure and arrival delays. For random forests, it has been found that a longer forecast horizon helps to increase accuracy while minimizing forecast error. Etani J. Big Data [8] Employing flight and meteorological data, a supervised model of on-time arrival flights is used. Peach Aviation's pressure patterns and flight data were discovered to be related. The use of a random forest as a classifier allows for a 77% accurate prediction of flights arriving on time.

III. PROPOSED METHODOLOGY

A. Dataset

The dataset contains information on airlines, airports, and flights for the year 2015. There are three CSV files in the dataset, namely airlines, airports, and flights.

The Airlines file contains information about the airlines. It has two columns, IATA_CODE, and AIRLINE, with a dataset size of 359 bytes. The IATA_CODE is the three-letter code assigned to each airline by the International Air Transport Association (IATA), while the AIRLINE column contains the name of the airline.

The Airports file contains information about the airports. It has seven columns, IATA_CODE, AIRPORT, CITY, STATE, COUNTRY, LATITUDE, and LONGITUDE, with a dataset size of about 23KB. The IATA_CODE is the three-letter code assigned to each airport by the IATA. The AIRPORT column contains the name of the airport, CITY contains the name of the city where the airport is located, STATE contains the name of the state where the airport is located, and COUNTRY contains the name of the country where the airport is located. The LATITUDE and

LONGITUDE columns contain the geographical coordinates of the airport.

The Flights file contains information about the flights. It has 31 columns and a dataset size of about 54 MB. The columns include YEAR, MONTH, DAY, DAY_OF_WEEK, AIRLINE, FLIGHT_NUMBER, TAIL_NUMBER, ORIGIN_AIRPORT, DESTINATION_AIRPORT, SCHEDULED_DEPARTURE, DEPARTURE_TIME, DEPARTURE_DELAY, TAXI_OUT, WHEELS_OFF, SCHEDULED_TIME, ELAPSED_TIME, AIR_TIME, DISTANCE, WHEELS_ON, TAXI_IN, SCHEDULED_ARRIVAL, AIRLINE_DELAY, ARRIVAL_TIME, ARRIVAL_DELAY, DIVERTED, CANCELLED, CANCELLATION_REASON, AIR_SYSTEM_DELAY, SECURITY_DELAY, LATE_AIRCRAFT_DELAY, and WEATHER_DELAY.

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways

	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
0	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.44040
1	ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.68190
2	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
3	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
4	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT
0	2015	1	1	4	AS	98	N407AS	ANC	SEA
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI
2	2015	1	1	4	US	840	N171US	SFO	CLT
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA
4	2015	1	1	4	AS	135	N527AS	SEA	ANC

5 rows × 31 columns

Fig. 1 Snapshot of the first 5 rows in the dataset

The Flights file is the most extensive dataset in the given dataset, containing information on flights. The columns contain information such as the airline code, flight number, tail number, origin and destination airports, scheduled departure and arrival times, actual departure and arrival times, departure and arrival delays, and reasons for cancellation, among others. The given dataset can be used for various data analysis tasks such as understanding the performance of airlines, analyzing the reasons for delays and cancellations, identifying trends and patterns, and predicting flight delays, among others. The dataset can also be used for developing machine learning models to predict flight delays and improve airline operations.

B. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is a method of analyzing and visualizing data in order to summarize its main characteristics and gain insights into its underlying structure. The goal of EDA is to uncover patterns and relationships in the data that may not be immediately obvious. In the given dataset, the airlines file contains 14 rows and 2 columns, the airports file has 322 rows and 7 columns, and the flights file has 545571 rows and 31 columns.

```
Number of distinct airports: 322
Number of distinct airlines: 14
Number of distinct flights: 6345
```

Fig.2 Distinct count in the dataset

The insights found on the count of distinct airports, airlines, and flights were shown in the above figure 2.

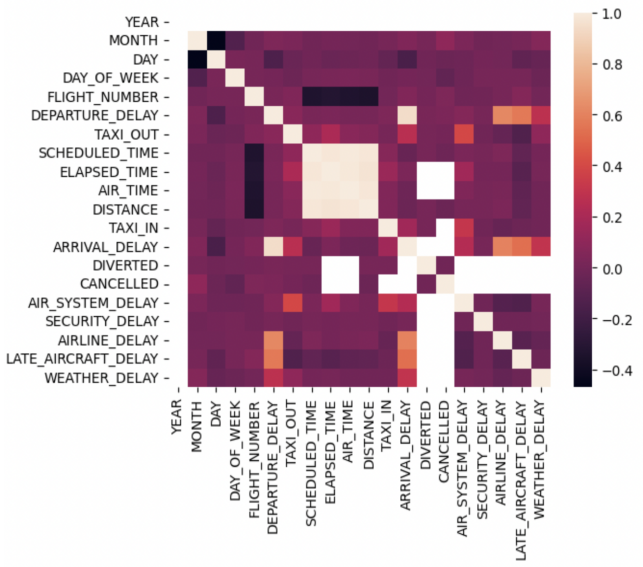

Fig. 3 A heatmap of flight data

A heatmap is a graphical representation of data that uses color-coding to represent values in a matrix or table. Heatmaps are useful for visualizing complex data sets, especially those with a large number of variables or categories. In a heatmap, each row and column of the matrix is assigned a color based on its numerical value. Typically, a gradient of colors is used to represent the range of values in the matrix, with darker colors indicating higher values and lighter colors indicating lower values. The color scale used in a heatmap can be customized to suit the specific needs of the analysis.

C. Data pre-processing/data wrangling

Data wrangling is an important step in the data analysis process, as it ensures that the data is accurate, consistent, and in a format that can be easily analyzed. Data wrangling can be a time-consuming process, but it is essential for ensuring that the results of the analysis are reliable and meaningful. We must first carry out fundamental pre-processing on our data set before we can apply algorithms to it. Since real-world data is irregular,

noisy and incomplete. Preprocessing is done to make the data more usable for analysis as well as to enhance its quality. The Bureau of Transportation has provided us with a set of data for 2015. 545571 rows and 31 columns make up the flight data set.

```
Number of null values in each column
 IATA_CODE    0
AIRLINE      0
dtype: int64

Number of null values in each column
 IATA_CODE    0
AIRPORT      0
CITY         0
STATE        0
COUNTRY      0
LATITUDE     3
LONGITUDE    3
dtype: int64

Number of null values in each column
 YEAR                      0
MONTH                      0
DAY                        0
DAY_OF_WEEK                0
AIRLINE                    0
FLIGHT_NUMBER              0
TAIL_NUMBER             3881
ORIGIN_AIRPORT             0
DESTINATION_AIRPORT        0
SCHEDULED_DEPARTURE        0
DEPARTURE_TIME         17545
DEPARTURE_DELAY        17545
TAXI_OUT               17855
WHEELS_OFF             17855
SCHEDULED_TIME             1
ELAPSED_TIME           19210
AIR_TIME               19210
DISTANCE                   0
WHEELS_ON              18343
TAXI_IN                18343
SCHEDULED_ARRIVAL          0
ARRIVAL_TIME           18343
ARRIVAL_DELAY          19210
DIVERTED                   0
CANCELLED                  0
CANCELLATION_REASON    527586
AIR_SYSTEM_DELAY       432210
SECURITY_DELAY         432210
AIRLINE_DELAY          432210
LATE_AIRCRAFT_DELAY    432210
WEATHER_DELAY          432210
dtype: int64
```

Fig.4 Null values in the dataset

There were numerous rows with blank or empty values. The data set was cleaned up by removing the rows and columns that contained null values using the dropna() function of the Pandas programming language.

```
YEAR                    0
MONTH                   0
DAY                     0
DAY_OF_WEEK             0
AIRLINE                 0
FLIGHT_NUMBER           0
TAIL_NUMBER             0
ORIGIN_AIRPORT          0
DESTINATION_AIRPORT     0
SCHEDULED_DEPARTURE     0
DEPARTURE_TIME          0
DEPARTURE_DELAY         0
TAXI_OUT                0
WHEELS_OFF              0
SCHEDULED_TIME          0
ELAPSED_TIME            0
AIR_TIME                0
DISTANCE                0
WHEELS_ON               0
TAXI_IN                 0
SCHEDULED_ARRIVAL       0
ARRIVAL_TIME            0
ARRIVAL_DELAY           0
DIVERTED                0
CANCELLED               0
CANCELLATION_REASON     0
AIR_SYSTEM_DELAY        0
SECURITY_DELAY          0
AIRLINE_DELAY           0
LATE_AIRCRAFT_DELAY     0
WEATHER_DELAY           0
dtype: int64
```

Fig. 5 Removal of nulls

D. Data processing

A histogram is a type of graph that displays the distribution of numerical data. In a histogram, the data is divided into a set of intervals, or "bins", and the frequency or count of data points that fall into each bin is represented by the height of a bar. The width of each bin is usually equal, and the bins are often chosen so that they cover the range of values in the data.

Histograms are particularly useful when dealing with large datasets, as they provide a visual representation of the distribution of the data, making it easy to see patterns and outliers.

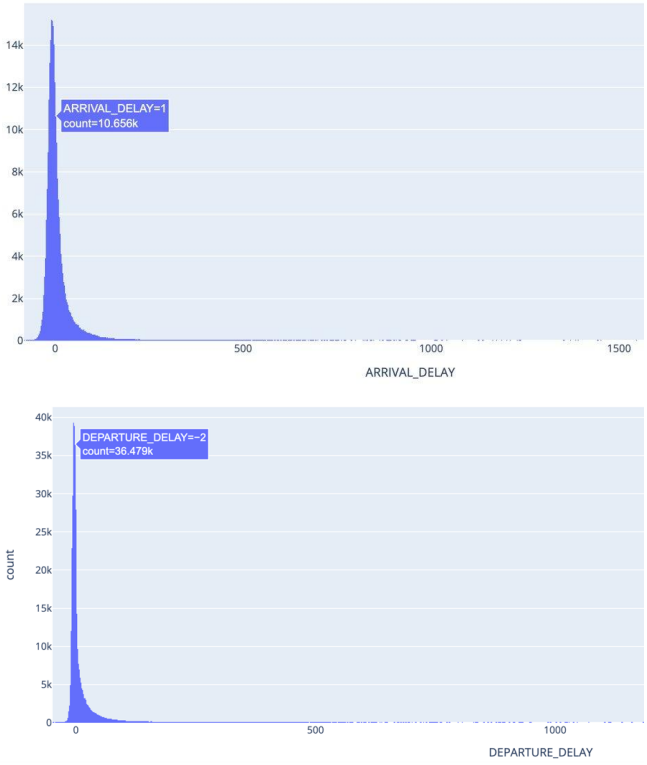

Fig. 6 Histogram of Departure and Arrival Delay

In Python, the describe() function is a method of Pandas DataFrame and Series objects that computes various summary statistics of the data. The output of describe() can provide a quick summary of the data, including its central tendency, variability, and distribution, and can be useful in analyzing the data.

```
flights.describe()[['ARRIVAL_DELAY', 'DEPARTURE_DELAY']]
```

	ARRIVAL_DELAY	DEPARTURE_DELAY
count	526361.000000	526361.000000
mean	6.532655	10.217767
std	40.642087	37.590598
min	-82.000000	-48.000000
25%	-12.000000	-5.000000
50%	-3.000000	-1.000000
75%	11.000000	9.000000
max	1971.000000	1988.000000

Fig.7 Flight data description

From Figure 7, it can be noted that count is the number of non-missing values in the data, mean is the average of the data, std is the standard deviation of the data, min is the minimum value in the data, 25% is the 25th percentile value of the data, 50% is the 50th percentile (median) value of the data, 75% is the 75th percentile value of the data, and max is the maximum value in the data.

A judgment on the presence of outliers can be made after studying the above data description. A box plot, also known as a box-and-whisker plot, are a graphical representation of the distribution of numerical data based on five summary statistics: the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value. The box in the middle of the plot represents the interquartile range (IQR), which is the difference between Q3 and Q1. The median is typically shown as a line inside the box. The whiskers, which extend from the box, represent the range of data outside the IQR, with any points outside the whiskers considered outliers. In a box plot, outliers are defined as any data point that falls outside the range of 1.5 times the IQR from either Q1 or Q3.

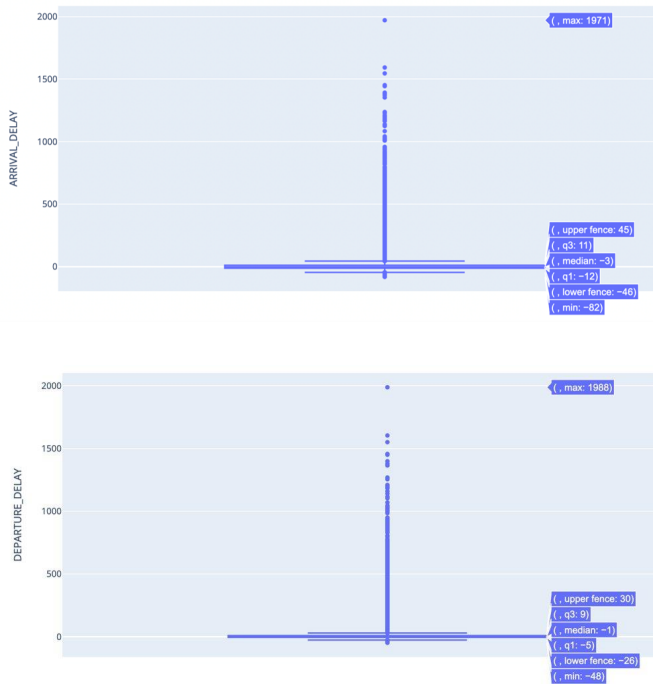

Fig. 8 Box plot of arrival and departure delays

A significant portion of the above figure is abnormal. The minimum values are -81 and -48 minutes for arrival and departure delays, respectively. The maximum values are 1971 and 1988 minutes for arrival and departure delays, respectively. Without taking appropriate action on the outliers, further analysis should not be done because the results may differ.

E. 3-Sigma rule

The 3-sigma rule, also known as the empirical rule, is a statistical guideline that states that in a normal distribution, approximately 68% of the data falls within one standard deviation of the mean, approximately 95% falls within two standard deviations of the mean, and approximately 99.7% falls within three standard deviations of the mean.

The 3-sigma rule is a useful tool for identifying outliers, as any data points that fall outside the range of three standard deviations from the mean are considered rare events with a probability of less than 0.3%. These data points may indicate unusual or unexpected behavior in the data and warrant further investigation. It is important to note that the

3-sigma rule, applies specifically to normal distributions and may not be applicable to other types of distributions. Additionally, the rule is based on an approximation and should not be used as a strict rule for making decisions.

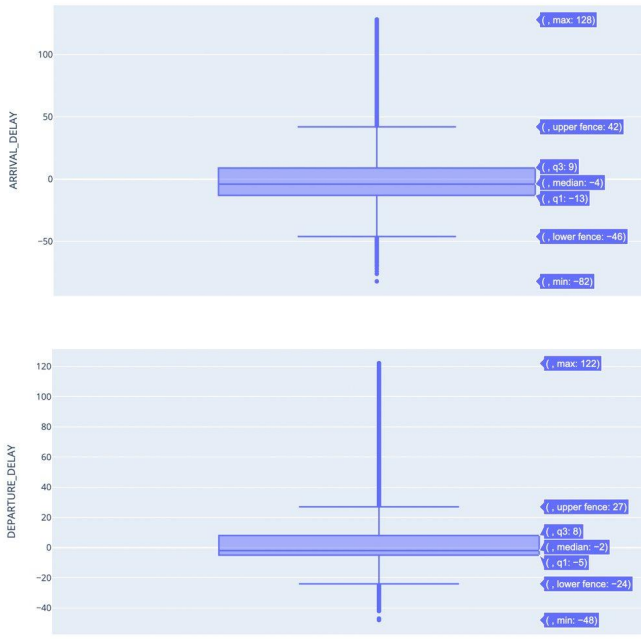

Fig. 9 Box plot after 3-sigma rule application

The maximum values are 128 and 122 minutes for arrival and departure delays, respectively. The outliers were replaced with mean*3 (standard deviation) values.

F. Data Visualization

Data visualization is the graphical representation of data and information. It is a powerful tool for analyzing and communicating complex data, as it allows us to visually explore patterns, trends, and relationships that may not be immediately apparent from raw data.

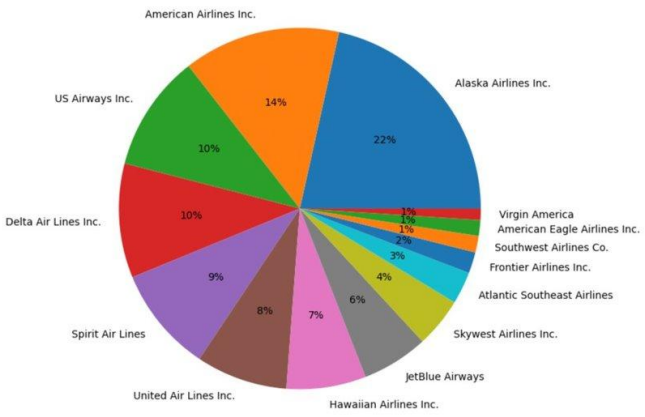

Fig.10 Piechart showing the percentage contribution of flights by each airline

From Figure 10, it can be noted that Alaska Airlines runs more flights, followed by US Airways and Delta Air Lines. While Virgin America and American Eagle Airlines contribute the least to the total.

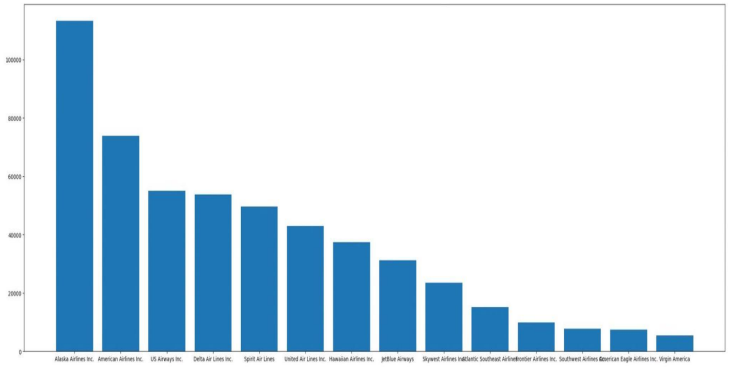

Fig. 11 A bar chart showing the number of flights operated by each airline

A bar chart showing the number of flights operated by each airline is a graphical representation of the frequency of flights operated by different airlines. The chart consists of a horizontal axis that represents the airlines, and a vertical axis that represents the number of flights operated. Each airline is represented by a rectangular bar, with the height of the bar corresponding to the number of flights operated by that airline.

IV. PROPOSED MACHINE LEARNING MODELS

A. Linear Regression Model

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the dependent variable and the independent variable(s), which means that changes in the independent variable(s) are associated with a proportional change in the dependent variable.

In simple linear regression, there is only one independent variable, and the relationship between the independent variable and the dependent variable is modeled using a straight line. In multiple linear regression, there are multiple independent variables, and the relationship between the independent variables and the dependent variable is modeled using a linear equation.

```
Data columns (total 16 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   YEAR                 100000 non-null  int64
 1   MONTH                100000 non-null  int64
 2   DAY                  100000 non-null  int64
 3   LATE_AIRCRAFT_DELAY  100000 non-null  float64
 4   AIRLINE_DELAY        100000 non-null  float64
 5   AIR_SYSTEM_DELAY     100000 non-null  float64
 6   WEATHER_DELAY        100000 non-null  float64
 7   ELAPSED_TIME         100000 non-null  float64
 8   DEPARTURE_DELAY      100000 non-null  float64
 9   SCHEDULED_TIME       100000 non-null  float64
 10  AIR_TIME             100000 non-null  float64
 11  DISTANCE             100000 non-null  int64
 12  TAXI_IN              100000 non-null  float64
 13  TAXI_OUT             100000 non-null  float64
 14  DAY_OF_WEEK          100000 non-null  int64
 15  SECURITY_DELAY       100000 non-null  float64
```

The code first selects a set of input features (called ip_feat) that are likely to be relevant in predicting flight delays. These features include things like the

'YEAR','MONTH','DAY','LATE_AIRCRAFT_DELAY','AIRLINE_DELAY','AIR_SYSTEM_DELAY','WEATHER_DELAY', 'ELAPSED_TIME', 'DEPARTURE_DELAY', 'SCHEDULED_TIME','AIR_TIME','DISTANCE','TAXI_IN', 'TAXI_OUT', 'DAY_OF_WEEK', 'SECURITY_DELAY'.

Next, a random subset of 100,000 flights is selected from the larger dataset (flights_new). The input features are stored in a variable called input_flights_data, and the output variable (whether the flight was delayed or not) is stored in a variable called output_flights_data.

```
model=LinearRegression()
model=model.fit(input_flights_data_train,output_flights_data_train)
slope=model.coef_
coef=model.intercept_
print(slope.flatten())
print(coef)

predictions = model.predict(input_flights_data_train)

r2_score(output_flights_data_train,predictions)
```

Fig. 12 The sample code of Linear Regression Model

The dataset is then split into training and testing sets using the train_test_split function, with 20% of the data used for testing. The scale function is used to standardize the input data by removing the mean and scaling to unit variance.A linear regression model is then created using the LinearRegression function from scikit-learn. The fit method is called on the training data to train the model. The model is then used to make predictions on the training data, and the R-squared score is calculated using the r2_score function to evaluate the model's performance.

R^2 Score: In regression analysis, the R-squared (R^2) score, also known as the coefficient of determination, is a statistical measure that represents the proportion of variance in the dependent variable (y) that can be explained by the independent variable(s) (x) included in the regression model.

The R-squared score ranges from 0 to 1, with a higher value indicating a better fit of the regression model to the data. An R-squared score of 1 indicates that the regression model perfectly fits the data, while a score of 0 indicates that the model does not explain any of the variance in the dependent variable.

The R-squared score is calculated by comparing the sum of the squared differences between the actual values of the dependent variable and the predicted values by the regression model (sum of squares residual, SSR) to the total sum of squares (SST).

The formula for calculating R-squared is:

$$R - Squared = 1 - \frac{First\ Sum\ of\ Errors}{Second\ Sum\ of\ Errors}$$

```
[ 0.         -0.01114832 -0.01736648 -0.62061036 -0.75608612 -0.36494152
 -0.30902518  1.09600602  1.28179191 -2.1681983   1.0894752  -0.05408788
  0.11687283  0.1917817   0.0027649  -0.01763839]
0.40665
0.6053897373005964
```

Fig. 13 The R^2 Score of the Linear Regression Model

After performing a linear regression model on the data set, its R^2 score is 0.605, which shows a chance of improvement and accurate results.

B. Model Selection

A code has been written to check which model suits the dataset better to perform further analysis to get more accurate results and a better R^2 score. It performs a comparison of multiple machine learning regression models using the cross_val_score() method and stores the evaluation metrics in a list called output_res.

	Model	Train Acc	Test Acc	AUC
0	(DecisionTreeRegressor(random_state=1939384561...	0.9677	0.8326	0.988378
1	(DecisionTreeRegressor(max_depth=3, random_sta...	0.5146	0.5178	0.868465
2	([DecisionTreeRegressor(criterion='friedman_ms...	0.6616	0.6628	0.962683
3	(ExtraTreeRegressor(random_state=1033062342), ...	1.0000	0.7464	0.978835
4	KNeighborsRegressor()	0.7095	0.5574	0.916162
5	(DecisionTreeRegressor(max_features=1.0, rando...	0.9785	0.8538	0.993617
6	ExtraTreeRegressor()	1.0000	0.3835	0.845895
7	DecisionTreeRegressor()	1.0000	0.6803	0.919497

Fig. 14 Comparison of Different Models

From the output provided, the Random Forest Regressor has the highest test accuracy score (0.8511) and the area under the ROC curve (AUC) of the model's predictions on the test data (AUC value) (0.9931), which suggests it may be the best model for this specific dataset and task.

C. Random Forest Regression Model

Random forest regression is a machine learning algorithm used for regression tasks. It is an extension of the Random Forest algorithm used for classification tasks.

Random forest regression builds multiple decision trees by randomly selecting a subset of features from the input data and a subset of samples from the training data. The algorithm constructs each decision tree using the selected features and samples, and the final prediction is an average of the predictions made by all the trees.

Random forest regression has several advantages over other regression algorithms. It can handle large datasets, it is resistant to overfitting, and it can capture complex nonlinear relationships between the input features and the output variable.

```
randomforest_model = RandomForestRegressor(n_estimators = 100, random_state=42)

output = flights_new['ARRIVAL_DELAY']
output = np.array(output)

input_flights_data = np.array(input_flights_data)

input_train, input_val, output_train, output_val = train_test_split(input_flights_data,

randomforest_model.fit(input_train, output_train)

late_pred = randomforest_model.predict(input_val)

print('Mean Absolute Error(MAE): ', round(np.mean(abs(late_pred - output_val)),3), 'min
```

```
Mean Absolute Error(MAE):  1.608 minutes.
```

Fig. 15 Sample code of the Random Forest Regressor Model

The code above implements a random forest regression model using scikit-learn's RandomForestRegressor class. The model is trained on the input and output features from the flight dataset to predict the arrival delay of a flight.

The input features used to train the model are split into training and validation sets using the train_test_split function. The model is trained using the training set and the fit method of the RandomForestRegressor class. Once the model is trained, it is used to make predictions on the

validation set using the predict method. The mean absolute error (MAE) of the predicted arrival delay values and the actual arrival delay values is calculated using Numpy's mean and abs functions.

Overall, the code is evaluating the performance of a random forest regression model in predicting the arrival delay of flights using the MAE metric.

Mean Absolute Error: It is a common metric used to evaluate the performance of regression models. The MAE measures the average absolute difference between the predicted and actual values in a set of predictions. The smaller the MAE, the better the model is performing. It is calculated by taking the average of the absolute differences between the predicted and actual values:

$$MAE = (1/n) * \Sigma|i=1 \text{ to } n|(|y\_pred,i - y\_true,i|)$$

where y_pred is the predicted label, y_true is the true label, and n is the number of samples used. From Figure 15, we can note that the MAE using the random forest regressor method is 1.608 minutes.

```
input_flights_data_df=pd.DataFrame(input_flights_data)

imp_features=randomforest_model.feature_importances_
imp_features=pd.DataFrame([input_flights_data_df.columns,imp_features]).transpose()
imp_features.columns=[['Variables','Importance']]
imp_features

print("r^2 score of training data: ",randomforest_model.score(input_train, output_train))
print("r^2 score  of test data: ", randomforest_model.score(input_val, output_val))
```

```
r^2 score of training data:  0.9982880040178168
r^2 score  of test data:  0.9921842872228424
```

Fig. 16 The R^2 Score of the Random Forest Regression Model

From the above figure, it can be noted that the R^2 score of both training and test data was approximately 1.

```
MAE for initial data sample:  1.6077 minutes.
MAE for current data sample:  1.4816 minutes.
r-squared score for initial data sample:  0.9928509155681301
/usr/local/lib/python3.9/dist-packages/sklearn/base.py:432: UserWarning:

X has feature names, but RandomForestRegressor was fitted without feature names

r-squared score for current data sample:  0.9922933409245321
```

Fig. 17 The R^2 Score of the Current Sample

From the above figure, it can be noted that the R^2 score of a data sample was 0.9922, approximately 1, which shows that the performance of the model is accurate.

V. CASE STUDIES

```
#Based on Airlines
flights['AIRLINE'].value_counts()
airline_name = "Southwest Airlines Co."
particular_airline = flights[flights['AIRLINE'] == airline_name]
i = np.random.randint(0, len(particular_airline))

particular_airline.iloc[i]

input_f = particular_airline.loc[:,ip_feat]
input_f = input_f.iloc[i]
out_y =particular_airline.iloc[i]['ARRIVAL_DELAY']
print("out_y: ", out_y)
input_f

predicted_val = randomforest_model.predict([input_f])
print("Delay predicted in mins: ", predicted_val)
print("Actual flight Delay in mins:  ", out_y)
print("Difference in prediction & actual delay: ",  out_y-predicted_val)
```

```
out_y:  6.0
Delay predicted in mins:  [5.44]
Actual flight Delay in mins:   6.0
Difference in prediction & actual delay:  [0.56]
```

Fig.18 Case Study Based on Airline

This code chooses a random record of a flight from the dataset that is associated with a specific airline, in this case "Southwest Airlines Co." The arrival delay for that flight is predicted using the input features in "input_f" by the trained random forest regression model ("randomforest_model"). "predicted_val" contains the anticipated delay information.

The projected delay is 5.44 minutes, the actual delay is 6.0 minutes, and the difference between the predicted delay and the actual delay is 0.56 minutes.

```
#Based on Flight Number
flights['FLIGHT_NUMBER'].value_counts()
# flight_number= 127
def flight_function(flight_number):
    try:
        particular_flight = flights[flights['FLIGHT_NUMBER'] == flight_number]
        i = np.random.randint(0, len(particular_flight))
        particular_flight.iloc[i]
        input_f = particular_flight.loc[:,ip_feat]
        input_f = input_f.iloc[i]
        out_y =particular_flight.iloc[i]['ARRIVAL_DELAY']
        predicted_val = randomforest_model.predict([input_f])
        print("Delay predicted in mins: ", predicted_val)
        print("Actual flight Delay in mins:  ", out_y)
        print("Difference in prediction & actual delay: ",  out_y-predicted_val)
    except:
        print("Flight Number Not found")

flight_function(98)
```

```
Delay predicted in mins:  [0.83]
Actual flight Delay in mins:   1.0
Difference in prediction & actual delay:  [0.17]
```

Fig.19 Case Study Based on Flight Number

This code chooses a random record of a flight, in this case "98". The arrival delay for that flight is predicted using the input features in "flight_function" by the trained random forest regression model ("randomforest_model").

The projected delay is 0.83 minutes, the actual delay is 1.0 minutes, and the difference between the predicted delay and the actual delay is 0.17 minutes.

VI. CONCLUSION AND FUTURE WORKS

From the results of the above case studies (Figs. 18 and 19), we can conclude that the random forest regression model works well for prediction with very high accuracy.

The future scope of this paper is to find the accuracy of the model by performing regression on the outliers to get more accurate and exact results.

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

[6] 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.

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

[8] Noriko, Etani, "Development of a predictive model for on-time arrival
fight of airliner