Airline Delay Predictions using Supervised Machine Learning

PranalliChandraa and Prabakaran.N and Kannadasan.R, VIT University, Vellore. pranalli24@gmail.com, dhoni.praba@gmail.com and desurkannadasanr@gmail.com

Abstract—The primary goal of this project is to predict airline delays caused by various factors. Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions. Hence, these factors indicate how necessary and relevant it has become to predict the delays no matter the wide-range of airline meshes. 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 3. This prediction will be helpful for giving a detailed analysis of the performance of individual airlines, airports, and then making a well-assessed decision. Moreover, apart from the assessment related to the passengers, delay prediction analysis will also help in important decision-making procedures necessary for every pivotal player in the air transportation system.

I. INTRODUCTION

During the most defining period of human history, where computing has moved from mainframes to PCs to cloud, and now to artificial intelligence. A fundamental sub-area of artificial intelligence has come into notice, called as Machine Learning, which enables computers to get into a mode of self-learning without being explicitly programmed. With the concept of machine learning, we have been able to apply complex mathematical computations to big data iteratively and automatically, that too with efficient speed, this phenomenon has been encompassing momentum over the last several years. On the other hand, data mining involves data discovery and sorting it among large data sets available to identify the required patterns and establish relationships with the aim of solving problems through data analysis. Simply combining, machine learning and data mining use the same type of approach and set of algorithms, except the kind of data pre-processing and end prediction varies. BY combining these two core areas to predict and present the most accurate results possible.

A. Supervised Machine Learning

It is a machine learning task where the dataset inputs and outputs are clearly recognized and already given, then several type of algorithms are trained using labeled examples. A supervised learning algorithm contains an entire dataset, which is further divided into training and test data; the algorithm examines the training dataset and produces an inferred function, which is then used for mapping new examples. In case of the aviation industry, commercialized aviation is a type of transportation system that is complexly distributed. It tends to deal with several important resources, demand fluctuations, and various other kinds of stages. Stages are bound to take place at terminal boundaries, runways, airports, and distinguished airspaces that may be susceptible to different kind of delays or cancellations. Summing up, some set of examples include weather conditions, ground delays, air traffic control and several other constraints and unforeseen circumstances that lead to delays and cancellations in the entire aviation industry. Hence, this becomes an optimal scenario which will allow us to implement a supervised machine learning algorithm to precisely determine and predict the class labels for unrevealed instances.

Supervised Learning algorithm here will model relationships and dependencies between the aimed prediction output and the input features, such that I’ll be predicting the output values for new data based on the relationships which are learned from the previous data set. Supervised Learning problems can be further categorized into following problems

• **Classification** – It is a type problem in which the output variable is an entire category itself, such as “Win” or “Lose”, the entire input data is classified into the category variables; it is generally used largely for recommendation problems

• **Regression** – It is a type of problem is which the output variable is a real value, such as few raw data values related to something. This is the problem type massively used for prediction analysis, and hence will be used in this project.

B. Regression Analysis Methods

The main focus of regression analysis is to model and determine the expected value of a dependent variable \( y \) in terms of the value of one or more independent variables \( x \).

• **Linear Regression**

Linear Regression is used to model and establish a relationship between dependent and independent set of variables by fitting the best line possible. The best fit line
hence formed as the result of prediction carried out is known as our regression line and is represented by a linear equation (1):

\[ y = b_0 + b_1 x_1 \]  

(1)

In case of logistic regression, which is very much compared to linear regression, the outcome (dependent variable) has only limited number of discrete possible values. Whereas, linear regression analysis is the first best-suited method because it results in any one among the range of an infinite number of possible values.

- **Polynomial Regression**

  In practice, rather than performing a simple linear regression, we can improve the model doing a fit with a polynomial of order N, because, in many situations, such a linear regression model may not hold true, or even if it does, the accuracy is decreased. Doing so, it is necessary to define the degree N which is optimal to represent the data. Hence, here it is where polynomial regression analysis becomes the next best-suited method for the prediction analysis.

  Represented by equation (2):

\[ y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n \]  

(2)

- **Multiple Linear Regression**

  If set of variables have a linear relationship with the dependent variable, then the regression is known as multiple linear regression. A multiple regression is represented by the following equation (3):

\[ y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \]  

(3)

In all three equations above, (1), (2) and (3), \(b_0, b_1, \ldots, b_n\) are the coefficients of the equation whose values we need to determine in any model; the \(x_1, x_2, \ldots, x_n\) are the dependent variables involved; and \(y\) is the independent variable here.

Multiple Linear Regressions is an even wider class of regression that combines linear and non-linear regressions with multiple explanatory variables. In this case, because of the broad range of prediction possibilities it offers, using multiple regression in some of the models, which attempts to explain dependent variable using more than one independent variable.

**II. RELATED WORKS**

Flight Delays have become a common and complex phenomenon, it occurs due to the problems at the origin-airport, at the destination-airport, any ground reasons or a combination of these entire factors can also give rise to delays. Delays are also being regarded as caused due to specific airlines. Even if it is complex, it is still measurable with decent accuracy. And with respect to the schedule and on-time performance of airlines, their generally exists some pattern of flight delay (Wu, 2005)[4]. The results obtained from this project, Airline Delay Predictions using Supervised Machine Learning, it can help to better understand the phenomenon and up to a very large extent.

In 2013, it was estimated that approx. 36% of flights were delayed by more than five minutes in Europe, 32% of flights delayed by more than 15 minutes in the US, and 16% of flights were cancelled or sobered delays greater than 30-40 minutes in Brazil[1] Hence, it indicates how important this indicator is and how it acts no matter how wide the scale of airline meshes exists.

Furthermore, coming to the Indian scenario, in 2017, according to the reports by the Directorate General of Civil Aviation (DGCA), between January and April, close to 5.12 lakh domestic passengers in India faced issues due to airline companies denying boarding, as well as flight cancellations and delays [2]. Airline companies had to pay the passengers compensations of over Rs. 25 crore for various inconveniences during the first four months of this year. Hence, the prediction analysis retrieved from this project can contribute in the form of a prototype in helping to identify operational variables that contribute to delays in any country scenario.

(Allan et al., 2001)[3] analysed delays at NYC Airports from September ’96 through August’00, with the aim of finding out some major causes of delay occurred during the first year of an Integrated Terminal Weather System (or ITWS) use and delays occurred with ITWS in operation that were “avoidable” if in case weather conditions would have been improved. The methodology used in the study has considered some major causes of delays (for example, convective weather inside and outside the terminal area, and high winds), and these causes were generally neglected in previous studies of capacity constrained airports such as Newark International Airport (EWR). The research concluded that the usual methods of assessing delays only in terms of Instrument Meteorological Conditions (IMC), Visual Meteorological Conditions (VMC) and the respective airport capacities is way more simplified than required for determining the type of air traffic management investments that in the best ways reduces the possible “avoidable” delays.

330
Hansen and Hsiao, 2005) analysed the rise in flight delay in the United States domestic system by estimating an econometric model of average daily delay that combines the effects of arrival queuing, terminal weather conditions, seasonal effects, and secular effects (such as a half year). The results suggested that even after controlling these factors altogether, the delays decreased gradually from 2000 through mid-2003, but the trend reversed drastically thereafter.

(Rosen, 2002) measured the rate of change in flight timings that resulted due to infrastructure-constant changes in passenger demand. Results indicated that as the ratio of demand to fix infrastructure increased, the delays increased proportionately, which resulted in proper decrease in average flight times by approx. 7 minutes after the rapid decrease in the fall’01. The flight time differences between the airlines in the data sample were small, though the United Airlines had lesser average flight times in the winter quarter than America West, which is considered even smaller airline.

Over the past couple of years, various analytical models and simulation methods have been used to analyze flight delay, including deterministic queuing models, neural networks, econometric models etc. Although it is evident that the analysis on delays carried is either on macroscopic or microscopic data over a period of couple of days and this has happened because of the huge data of flights every day. Hence, the predictions led to less accurate results or relapse in the trend among the results. So here, obtaining the airline on-time performance data set from the U.S. DOT Bureau of Transportation Statistics (BTS) website, and the linear and polynomial regression models to be used along with regularization technique in machine learning is far better to identify the delay pattern. In this project, studies on airport delay and individual airlines delay behavior analysis are carried out, using linear regression model, polynomial regression models, and regularization. The performances of the models are tested using various metrics, e.g., CV Method, MSE/RMSE Scores, etc. This project will be able to complete several objectives like the statistical description of airlines, temporal variability of delays, the relation of delays with the origin airports, estimating geographically the flights from each airport, etc., along with the main prediction analysis.

III. REGRESSION ANALYSIS MODELLING

A. Overview of the Dataset

The dataset has been taken from a reliable online available government agency website that provides the air traffic delay statistics in the United States. The U.S. Department of Transportation’s (DOT) Bureau of TransportationStatistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. BTS compiles daily data for the benefit of the customers or for any data analysts. The dataset is of 2017 flight delays and cancellations.

<table>
<thead>
<tr>
<th>IATA_CODE</th>
<th>AIRLINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>UA United Air Lines Inc.</td>
</tr>
<tr>
<td>1</td>
<td>AA American Airlines Inc.</td>
</tr>
<tr>
<td>2</td>
<td>US US Airways Inc.</td>
</tr>
<tr>
<td>3</td>
<td>F9 Frontier Airlines Inc.</td>
</tr>
<tr>
<td>4</td>
<td>B6 JetBlue Airways</td>
</tr>
<tr>
<td>5</td>
<td>OO Skywest Airlines Inc.</td>
</tr>
<tr>
<td>6</td>
<td>AS Alaska Airlines Inc.</td>
</tr>
<tr>
<td>7</td>
<td>NK Spirit Air Lines</td>
</tr>
<tr>
<td>8</td>
<td>WN Southwest Airlines Co.</td>
</tr>
<tr>
<td>9</td>
<td>DL Delta Air Lines Inc.</td>
</tr>
<tr>
<td>10</td>
<td>EV Atlantic Southeast Airlines</td>
</tr>
<tr>
<td>11</td>
<td>HA Hawaiian Airlines Inc.</td>
</tr>
<tr>
<td>12</td>
<td>MQ American Eagle Airlines Inc.</td>
</tr>
<tr>
<td>13</td>
<td>VX Virgin America</td>
</tr>
</tbody>
</table>

Fig.2: All the airlines in the dataset associated with particular IATA carrier codes.

B. Data Exploration

Data cleaning is the critical initial step in evaluating the dataset for final analysis. With the enormous amount of data available, databases are prone to have noisy, missing and inconsistent data. The data in this project is obtained from...
BTS source, which has varying kinds of 31 variables involved, and may not be compatible with the format in which we require the data to use in Python. Data Cleaning helps in removing noisy data, and removing inconsistencies. Data cleaning is performed as follows:

**Dates and Times:**

The date format has been given in four variables format; it will be toned down to one particular format available in Python for ease of use.

**Filling Factor:**

In the data cleaning process, a missing value can be ignored, manually entered, given a constant value, or a mean value. In this case, it will be organizing and arranging the entire data frame to keep the relevant attributes and eliminate the ones which has missing values. This is done to increase the readability and feasibility of use.

The fill factor gives us what percentage of space on each page to fill with data. The fill factor value obtained, in general, can be defined as a percentage from 1 to 100. Here, it has been obtained a fill factor of >97%, which is quite satisfactory, that means 3% of the overall space can be used for future data growth.

Further, we have established statistical description of airlines, which involves classifying airlines on the basis of their punctuality; it is done using various statistical parameters

![Fig.3](image-url)

(a) Fig.3: (a) Pie chart with % of flights per company, (b) Mean delays of airlines at origin airports

![Fig.4](image-url)

Fig. 4: Comparative analysis of all the airlines with respect to their delays

Further, these are normalized the distribution of delays that modeled with an exponential distribution (Prabakaran, 2017) [8]:

$$F(x)=a e^{b}(-x)$$ \hspace{1cm} (4)

Both the parameters, a and b, have been obtained to describe each airline are given in the upper right corner of each panel in Fig.5(b). The normalization of the distribution implies that:

$$\int_{-\infty}^{\infty} F(x)dx = 1$$ \hspace{1cm} (5)

The normalization here implies to the histogram, and this relation entails that a and b coefficients will be correlated.
with \( a \propto \frac{1}{b} \) and hence, only one of these two values is necessary to describe the distributions. Finally, according to the value of either \( a \) or \( b \), a ranking of the airlines has been established: the low values of a will correspond to airlines with a large proportion of important delays and, on the contrary, airlines that beam from their punctuality will have high values.

Finally, according to the value of either \( a \) or \( b \), a ranking of the airlines has been established: the low values of \( a \) will correspond to airlines with a large proportion of important delays and, on the contrary, airlines that beam from their punctuality will have high values.

\[ \text{Fig. 5: (a)} \] Ranking of airlines based on delays
\[ \text{(b)} \] Individual graphs of airlines demonstrating \( a \) & \( b \) parameters

It is seen in the above Fig.5(a) that SouthWest Airlines, that represents \( \sim 20\% \) of the total number of flights, is ranked well and has occupied the third position. Furthermore, the arrival delay has been examined, and it is different from the departure delay, it is also retrieved that the arrival delay is not seen up to a very huge extent. Hence, only departure delay is considered.

\[ \text{C. Prediction of delays using regression} \]

It is deduced from various observations between the origin airports and all the airlines, that there is a high variability in average delays, both noticed between different airports but also between different airlines. This is significant cause it implies that in order to accurately model the delays, it is be necessary to adopt a model that is specific to the company and the home airport.

After the exploration of dataset, the final aim to achieve is to devise models for prediction of delays. The prediction is retrieved using a three week window that will predict the delays for the following week.

There are two models developed for the prediction of delays, which are as follows:

\[ \text{Model 1: One Airport – One Airline} \]

Here, delays are modeled by separately considering the airlines and by splitting the data according to the different home airports. This basic model is called as a "toy-model" that helps to identify problems that may arise at the production stage. It is to be made sure that the automation of the whole process is robust enough to insure the quality of the fits, which can occur while treating the whole data.

The pitfall that may occur is that of insufficient statistics or extreme delays. Extreme Delays are seen when the delay noted is extremely high (>10 hrs.), that may have occurred due to any unforeseen or unpredictable circumstance (e.g. weather conditions, accidents, etc.), this delay is rare and introduces a bias in the analysis. In conclusion, the way we handle delays determines and impacts the modeling to a large extent.

In practice, the model provides a better fit line with polynomial regression or order N. It is necessary to define the value N which provides the best results, and while increasing the N value, over-fitting needs to be avoided which happens when more data is been added to test and the model becomes even more complex, which in turn disrupts the local structure(over-fit). And this is avoided by splitting the datasets into test and training sets. The technique is made more robust by performing cross-validation method. This method consists of re-separating the data into test, training and validation sets. The learning is done on the training set, but to avoid over-learning this method facilitates split into several pieces that are used alternately for training and testing. The cross-validation method helps in avoiding any kind of bias in the estimation parameters because all of the data is used successively to drive the model.

The K-fold helps in choosing the best polynomial degree. It is seen in Fig.6 that the best model (best generalized model) is of order 2.

\[ \text{Fig. 6: Using the dataset, and applying K-fold method, the MSE values we get in Python 3.} \]

On this stage, after confirming the order of polynomial, as
it has been validated, the entire dataset is used in order to perform the fit. The following figure, Fig.7, compares the K=50 polynomial fits corresponding to the cross-validation calculation leads to the orange curve. The final model fit corresponds to the blue line.

MSE (Mean Square Error) value calculated for this model is 108.6713085. The MSE value gives us an idea of how close a regression fit line is to the original data points. It does this by taking the distances of the fitted line points from the data points (distance="errors") and summing the square of each of them. It finally takes the average of the value we get. The RMSE value (square root of MSE value) we get here is 10.42(min). It refers to the difference in minutes between the predicted delay and the actual delay, and in this case, the difference between the model and the observations found.

**Model 2: One Airline-All Airports**

In Model 1, only one airport was considered. This procedure is efficient only up to some extent because it is likely that some of the observations can be extrapolated from an airport to another. Thus, it is considered advantageous to make a single fit, which would take all the airports into account. Particularly, this would allow predicting delays on airports for which the number of data is low with a better accuracy.

Here, to test, it has been chosen as the carrier="AA", that is, American Airlines, and in the data frame, a label has been assigned to each airport. The correspondence between the label and the original identifier has been saved in a list Python. The next step involves incorporating the "One Hot Encoding" Method. In machine learning, to work with categorical variables, the categorical data is converted into numbers, which is required for both input and output data that are categorical. This method is applied in this case by creating a matrix where instead of the ORIGIN_AIRPORT variable that contained M labels, we build a matrix with M columns, filled with 1 and 0 depending on the correspondence with particular airports.

Linear Regression is first performed on this model, and extreme or large delays are underestimated and not taken into account, as explained. Fig.10 gives depicts this.

**Fig 7: Graph showing final fit and CV output get in Python 3.**

**Fig 8: MSE value and quality % of linear regression on Model 2 obtained in Python**

\[
\text{MSE} = 53.7490736042 \\
\text{5.30%}
\]

In practice, the quality of fit is also known by considering the number of predictions where the differences with data points (or real values) are greater than 15 minutes.

\[
\text{Quality} = \left( \frac{\text{No. of values >15min}}{\text{No. of predictions (total)}} \right) \times 100 \quad (6)
\]

The value found here is 5.30%.

Further, Polynomial Regression is performed on the fit, Fig.11 depicts it.

**Fig 9: MSE score and quality % of Polynomial fit in Model 2 obtained in Python**

\[
\text{MSE} = 49.502543214 \\
\text{4.81%}
\]

The MSE score found is 49.502543. The quality of the fit is again judged by the above formula, and is found 4.81%.

**Fig 10: Linear fit on Model 2.**
Hence, it is evident that a polynomial fit improves the MSE score slightly, and is an efficient model. Testing the model against end-week data, using regularization to minimize the errors and over fitting:

The current MSE score is calculated on the basis of all the airports that are served by American Airlines, whereas previously it was calculated on the data of a single airport. The current model is therefore more generalized and efficient.

IV. PERFORMANCE METRIC

**Cross Validation Technique and K-Fold Technique**

Cross Validation is a very important technique for assessing the performance of machine learning models. It enables us in knowing how a machine learning model would generalize to an independent data set.

The model dataset is divided into three sets: Training, test, and validation. The entire set is divided into K-folds or subsets, which is basically applying the K-fold technique, one of the ways of Cross Validation. Then, the K-1 folds are sent for training and the learning is done on it, then the model’s generalization is checked on the test set, which contains just the remaining one fold; and this process goes on till the last fold. This method is used in the initial stages of both Model 1 and Model 2, for data splitting and increased efficiency.

**MSE**

The Mean Squared Error (MSE) is a measure of how close a fitted line is to the real data points. For every data point on the line, we take the distance vertically from the real point to the corresponding Y value on the curve fitted (which is the error), and square the value. The next step is to carry out the summation of all the squared error values corresponding to all the data points, and, in the case of a linear fit, the value we get is divided by the total number of observations minus 2. The squaring is to avoid negative values cancelling the positive values. The quality of the model is assessed by the Mean Squared Error score we get, the smaller the value, the closer the fit is to the real data and the accurate the machine learning model.

\[
MSE = \frac{1}{n} \sum_{t=1}^{n} e_{t}^{2}
\]

(7)

\(e_t = \text{error value (predicted value-real value)}\)
\(n = \text{Total no. of attributes or points taken into account.}\)

MSE value for Model 1 is 108.6713085.
MSE value for Model 2, Linear Fit is (shown in Fig.8), 53.7430
MSE value for Model 2, Polynomial Fit (shown in Fig.12), 49.5025.

**RMSE**

Root Mean Squared Error (RMSE) is another quality that we calculate to measure the accuracy of a model. It is equal to the square root of the mean square error. It is considered as one of the most easily interpreted statistics, as it has the same units as the quantity plotted on the ordinate, which is the y-axis.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} e_{t}^{2}}
\]

(8)

\(e_t = \text{error value(predicted value-real value)}\)
\(n = \text{Total no. of attributes or points taken into account.}\)

The RMSE values are depicted using a variable Ecart, for both the models, in Fig.8 and Fig.9, as the least Ecart value shown is for Model 2, 4.81%.

As the MSE value and RMSE value is lowest for the polynomial regression on Model 2, hence, it depicts that it shows the most accurate results, and the fitted line (the predicted results) is the closest to the real data points.

Though, the final data set testing showed the Ecart value as 7.70 min. (Fig.12).
V. CONCLUSION

This project and the analysis retrieved are useful not only for passengers point of view, but for every decision maker in the aviation industry. Apart from the financial losses incurred by the industry, flight delay also portray a negative reputation of the airlines, and decreases their reliability. It causes various sustainability issues, for example, increase in fuel consumption and gas emissions. The analysis carried here not only predicts delays based on the previous available data, but also give statistical description of airlines, their rankings based on their on-time performance, and delays with respect to time, showing the peak hours of delay. This project can be used as a prototype by any aviation authority for their benefit, in the Indian Scenario too, it can work as an efficient model or a proper prototype to study delay analysis, based on the real dataset provided. This project has encompassed and showed the importance of Regression Analysis in Machine Learning, Data Mining concepts for efficient data cleaning, Cross Validation technique and Regularization in ML for making proper models and its predictive analysis.

REFERENCES
