**FLIGHT DELAY PREDICTION**

**STATISTICAL DATA ANALYSIS CA660 ASSIGNMENT**

**Student Names: Akanksha Rajpute, Joseph Cherian**

**Student Ids: 19210135, 19210222**

**Abstract:**

Flight delays are very prevalent around the world. Flight delays cost the industry nearly $25 billion per year. It not only affects the industry but also the civilians. The standard of the airline as a product has been devalued due to flight delays. While many air travellers choose to “grin and bear it” others choose to shift to alternative transportation modes. In the aviation industry, traveller’s decision on travelling and selection of airlines are based on the rate of flight delays[1]. Therefore, we have performed Logistic Regression and used Decision Trees for the prediction of flight delays. The objective of this analysis is to determine which flight would be delayed and by how long.

**Keywords** : Flights, Flight Delay, Prediction, Logistic Regression, Decision Trees, Regression

1. **Introduction**

As the world’s economy is moving forward, the demand for air travel is also increasing considerably. The issue of flight delay is getting serious, because of which the aviation industry is suffering from loses. For travellers, flight delay causes the trouble in travelling, bad mood, as well as the loss of time and economy; as for the airport, flight delay causes serious disruptions in the operation of the airport; and for an airline, they have to suffer economically and their reputation also degrades due to flight delays. Flight delay has become the shackles of the development of the aviation industry [3]. Flight delays lead to negative impacts, mainly economical for commuters, the airline industry and airport authorities. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions.

Regression is frequently used for prediction and forecasting. Linear and Logistic regression are popularly used for predictive analysis. Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables [2].

Decision Tree helps in making optimal decisions. It is well suited for problems where instances are represented by attribute value and when training data contains an error. It is also applicable to the situation when the target function has a discrete output value. It implicitly performs variable screening and requires relatively less effort from the user for data preparation [4].

In this project, we have applied logistic regression and decision trees to predict flight delays. We chose logistic regression over linear regression as it does not allow negative values and also helps in predicting as it has values between 0 to 1 only.

1. **Related Work**

[9] estimated flight departure delay distributions, and used the estimated delay information in a strategic departure delay prediction model. [3] Used multiple linear regression to predict flight delays.[10] Showed the effect on passengers because of flight delay.

[11] Proposed Bayesian Network to predict flight delays.

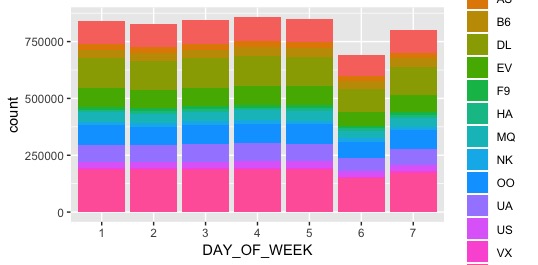
1. **Dataset and Exploratory Data analysis**

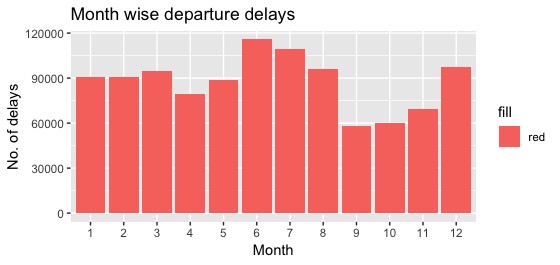
* The dataset contains 31 variables.
* Variable for Prediction: DEPARTURE\_DELAY.
* Delay greater than 15mins was considered as Delayed Flight.
* Missing or NA values are present for observations of Diverted & Cancelled Flights only. 1.8% of such observations are present
* Data: 14 Airlines, 628 Origin Airports & 629 Destination Airports.
* June and July are the months where more Departure Delay occurrences are Observed.
* Data Source: [12]

|  |  |
| --- | --- |
| **Field** | **Description** |
| YEAR | Year of the Flight Trip |
| MONTH | Month of the Flight Trip |
| DAY | Day of the Flight Trip |
| DAY\_OF\_WEEK | Day of the week of the Flight Trip |
| AIRLINE | Airline Identifier |
| FLIGHT\_NUMBER | Flight Identifier |
| TAIL\_NUMBER | Aircraft Identifier |
| ORIGIN\_AIRPORT | Starting Airport |
| DESTINATION\_AIRPORT | Destination Airport |
| SCHEDULED\_DEPARTURE | Planned Departure Time |
| DEPARTURE TIME | Actual Departure time |
| DEPARTURE\_DELAY | Total Delay on Departure |
| TAXI\_OUT | The time duration elapsed between departure from the origin airport gate and wheels off |
| WHEELS\_OFF | The time point that the aircraft's wheels leave the ground |
| SCHEDULED\_TIME | Planned time amount needed for the flight trip |
| ELAPSED\_TIME | AIR\_TIME+TAXI\_IN+TAXI\_OUT |
| AIR\_TIME | The time duration between wheels\_off and wheels\_on time |
| DISTANCE | Distance between two airports |
| WHEELS\_ON | The time point that the aircraft's wheels touch on the ground |
| TAXI\_IN | The time duration elapsed between wheels-on and gate arrival at the destination airport |
| SCHEDULED\_ARRIVAL | Planned arrival time |
| ARRIVAL\_TIME | WHEELS\_ON+TAXI\_IN |
| ARRIVAL\_DELAY | ARRIVAL\_TIME-SCHEDULED\_ARRIVAL |
| DIVERTED | Aircraft landed at the airport that was out of schedule |
| CANCELLED | Flight Cancelled (1 = cancelled) |
| CANCELLATION\_REASON | Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D – Security |
| AIR\_SYSTEM\_DELAY | Delay caused by air system |
| SECURITY\_DELAY | Delay caused by security |
| AIRLINE\_DELAY | Delay caused by the airline |
| LATE\_AIRCRAFT\_DELAY | Delay caused by aircraft |
| WEATHER\_DELAY | Delay caused by weather |

Table 1: Data fields and descriptions

**EDA Visualisation:**

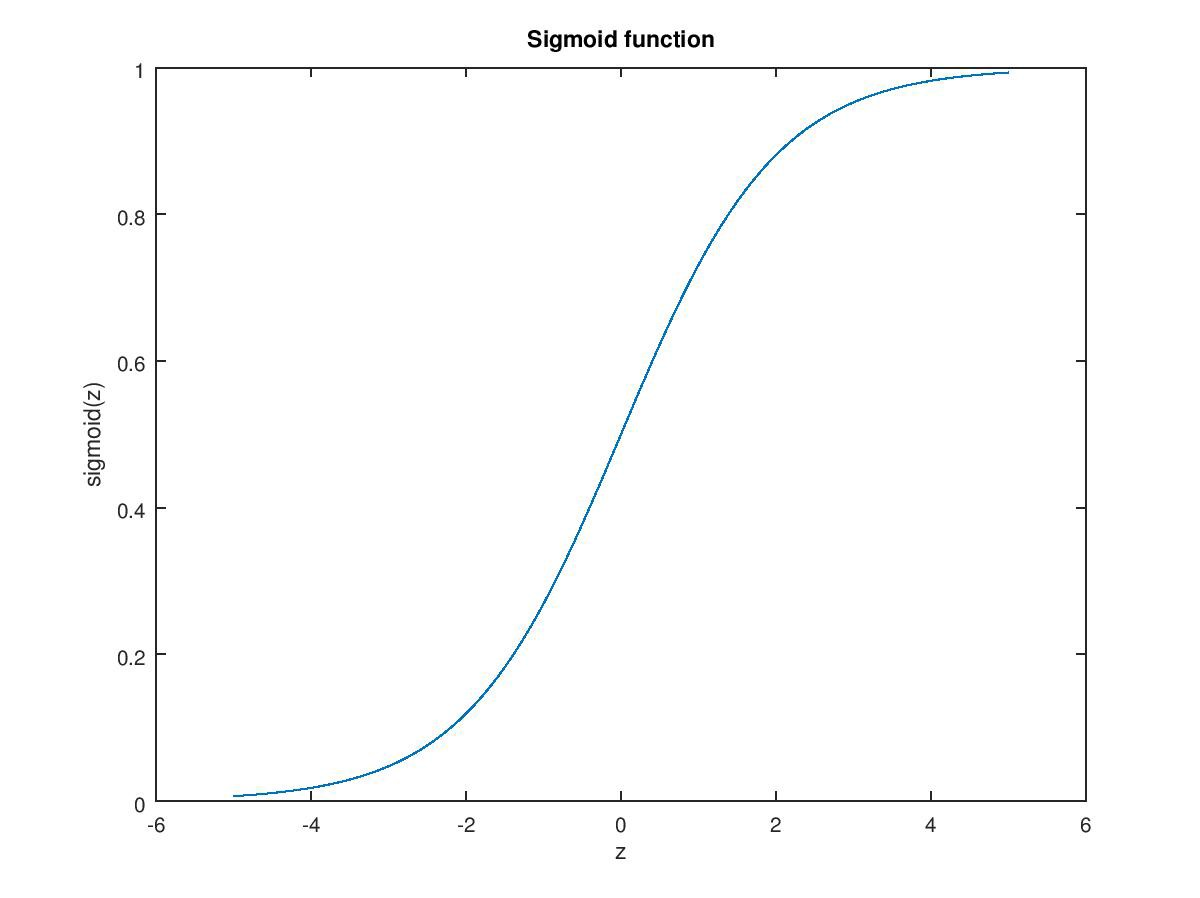
****



1. **Hypothesis and Research Questions**

* The primary goal of this project is to predict airline delays caused by various factors.
* Which method is better in predicting the flight delays?

1. **Method**
2. *Logistic Regression*

****Figure 1: The Logistic Curve g(Z) [5]

**Logistic regression** is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

In figure 1, g(z) = 1/( 1 + e ^−z), is called the l***ogistic function or the sigmoid function.***

In this project, we have used R to implement and predict the flight delay.

We have used some pre-existing code[5] and modified it according to our project as they have used logistic regression and random forest for prediction. We chose logistic regression as the predicted values lie between 0 and 1.

1. *Decision Trees*

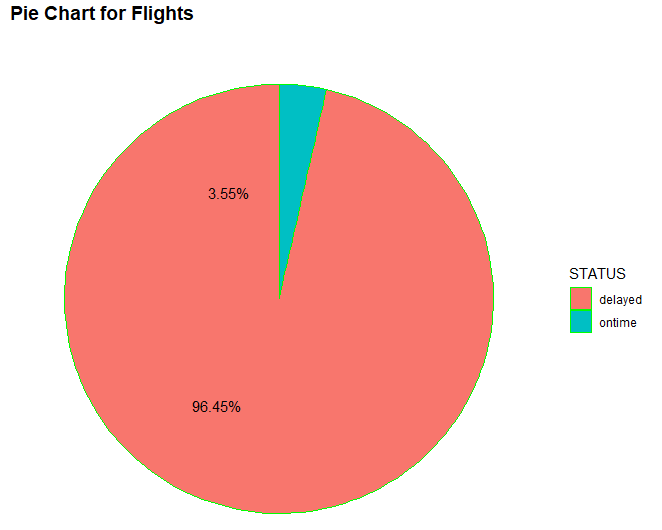
Decision trees can be used to solve Regression and Classification problems[8].

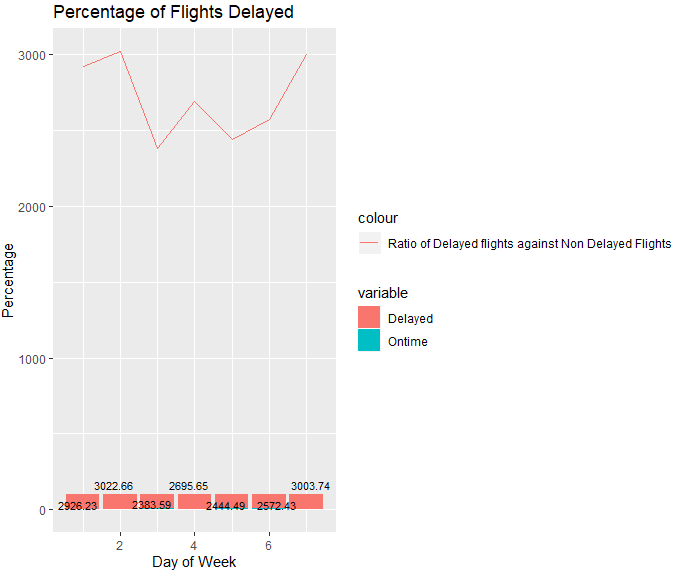
Decision Trees are predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question, and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface.

We have used the same project for implementation [5]. They have used Random Forest algorithm but instead, we used Decision Trees and modified that code according to our need.

1. **Results**





1. *Logistic Regression*

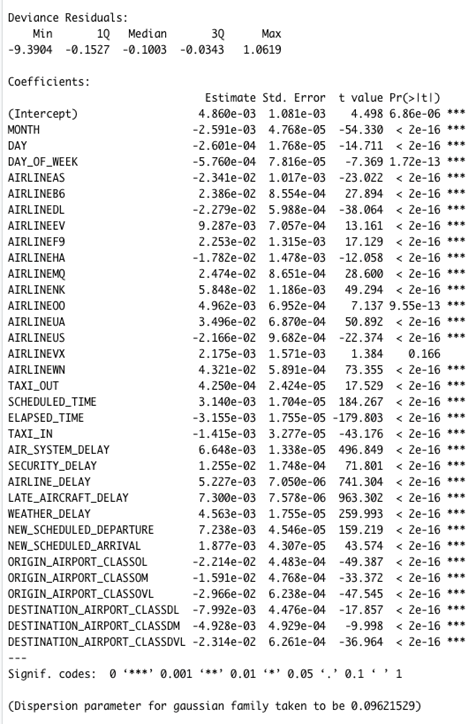
****

Figure 2: Logistic Regression Test Results

Trail1\_test Results : % Accuracy:87.5

1. *Decision Trees*

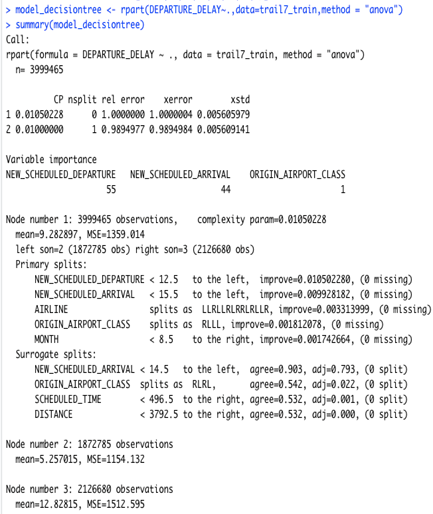
****

Figure 3: Decision Trees test results

Trail1\_test Results:

• RMSE: 36.75

1. **Conclusions**

We have observed that wind, temperature and other weather conditions play a major role in a flight delay. Also, Logistic regression is better for prediction in this scenario as the results shows. We have also observed that using Logistic Regression and Decision Trees altogether can yield better results.

References

[1] Ball, Michael; Barnhart, C.; Dresner, Martin; Hansen, Mark; Neels, K; Odoni, A.R.; Peterson, Everett; Sherry, Lance; Trani, Antonio; [Zou, Bo](https://rosap.ntl.bts.gov/gsearch?sm_creator=Zou%2C%20Bo),” Total delay impact study: a comprehensive

assessment of the costs and impacts of flight delay in the United States”,

University of California, Berkeley. Institute of Transportation

Studies(2010)

[2] Sunil Ray,”7 Regression Techniques you should know!

”,(2015)

[3] Yi Ding,” Predicting flight delay based on multiple linear regression ” IOP Conf. Ser.: Earth Environ. Sci. (2017)

[4] Vishakha Jha,” Decision Tree Algorithm for a Predictive Model

”,(2017)

[5] limcheekin, ”<https://github.com/limcheekin/r-flight-delay-prediction>”,(2016)

[6] Pramod Chandrayan, ”Logistic Regression For Dummies: A Detailed Explanation

”,(2015)

[7] J.S. Cramer,” The Origins of Logistic Regression”, *Tinbergen Institute,(2002)*

[8] [RahulSaxena](https://dataaspirant.com/author/rahul-saxena/), “HOW DECISION TREE ALGORITHM WORKS

”,(2017)

[9] Tu, Y., Ball, M.O., Jank, W.S., “Estimating flight departure delay distributions – a statistical approach with long-term trend and short-term pattern”, Am. Stat. Assoc. J., 103, pp. 112–125, (2008).

[10] Bratu, S., Barnhart, C. “An analysis of passenger delays using flight operations and passenger booking data”, Air Traffic Control Quart, 13 (1), pp. 1–27, (2005).

[11] Xu, N., Laskey, K.B., Donohue, G., Chen, C.H. “Estimation of delay propagation in the national aviation system using Bayesian networks”, In 6th USA/Europe Air Traffic Management Research and Development Seminar, (2005).

[12]<https://www.kaggle.com/usdot/flight-delays#flights.csv>