**Predicting Delay in Flights using Machine learning and Big Data**

**Project Report for CS-GY-6513 Big Data**

**A large airplane flying in the sky

Description automatically generated with low confidence**

***Submitted by***

**Remon Roshdy**

**rr3531**

**Table of Content**

|  |  |
| --- | --- |
| **Topics** | **Page Number** |
| **Purpose of Document** | **1** |
| **Problem statement and motivation** | **2** |
| **Dataset Description** | **3-5** |
| **Approach** | **5** |
| **Architecture**   * **Input Data** * **Apache Pyspark** * **Data Visualization** * **Machine Learning library** | **6** |
| **Exploratory Data analysis**   * **Missing Data** * **Duplicate Data** * **Relevant Columns** * **Columns with one value** * **Correlation Matrix** * **Once hot encoding** * **Sampling** | **7-10** |
| **Machine Learning models**   * **Logistic Regression** * **Random Forest** | **11-12** |
| **Reference** | **13** |

**Purpose:-**

This document details the implementation details carried out to analyze the flight arrival delays . The document describes the problem statement, approach, analysis, and the proposed architecture. The code and graphical representation of the analysis have been explained in detail for clarity.

**Problem Statement and Motivation**

Flight delay for Passengers causes inconvenience and loss of valuable time.

Flight delay for the airport affects the traditional operation of the airport

Flight delay for the airlines brings huge economic losses and also affects the reputation of the airline. When flight delay occurs it will affects the take off or landing of the next flight.

Decisions made by management of an airport are often very expensive and influence flight delay. Reducing flight delay helps to decrease costs and helps to improve the standard of the service provided to customers. For these reasons it is very important to search out which variables influence flight delays and use them to predict it.

We want to be able to predict, based on historical data, the arrival delay of a flight using only information available before the flight takes off. Since the target variable (ArrDelay - arrival delay) is a numerical value we will use regression algorithms and Random forest

**A picture containing diagram

Description automatically generated**

**Graphical user interface, text

Description automatically generated**

**Case study**

In this study, domestic flights data between Jan 2008 and Dec 2008 are used as a case

study. Note that the extracted dataset does not contain the international flights which have a relatively low delay ratio. The collected dataset, which is owned by The U.S.

Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) ,

contains a combined data of flights, origin of flight, destination, and distance, etc.

**Technology Used and Architectural**

This analysis leverages the following software applications:

# PySpark

* ML libraries

# Python & Library

* Numpy
* Pandas
* Scikit-learn

**A screenshot of a computer

Description automatically generated with low confidence**

**A picture containing text, room

Description automatically generatedLogo

Description automatically generated A picture containing text, clipart

Description automatically generated Logo, company name

Description automatically generated**

**The components used in the above architecture and their use is described below:**

1. Input datasource

Input data will be the flight data information for year 2008. We got the data from Harvard data source link above and ultimately data source coming from US department of transportation

1. Apache Spark

Spark is the heart of the processing in this architecture. Spark can handle large

Data set , our file has more than 2.3 Million rows.

1. Data Visualization

A tool like Plotly is used to show relevant information about the project.

**Dataset Description**

The results presented in this paper were obtained using data from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

We had access to a large dataset with 2.4 Million entries of data of commercial flights in the year 2008. The dataset provides information about flights by date of flight, the airport the flight will take off from, flight number, distance, departure, and arrival time.

**There are 29 fields in the dataset, and the main fields are below.**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Forbidden** | **Name** | **Description** |
| 1 |  | Year | 2008 |
| **2** |  | **Month** | **1-12** |
| **3** |  | **DayofMonth** | **1-31** |
| **4** |  | **DayofWeek** | **1-7** |
| **5** |  | **DepTime** | **Actual departure time** |
| **6** |  | **CRSDepTime** | **Scheduled departure time** |
| **7** | **X** | **ArrTime** | **Actual Arrival time** |
| **8** |  | **CRSArrTime** | **Scheduled Arrival time** |
| **9** |  | **UniqueCarrier** | **Unique carrier code** |
| **10** |  | **FlightNum** | **Flight number** |
| **11** |  | **TailNum** | **Plane tail number** |
| **12** | **x** | **ActualElapsedTime** | **In minutes** |
| **13** |  | **CRSElapsedTime** | **In minutes** |
| **14** | **X** | **AirTime** | **In minutes** |
| **15** |  | **ArrDelay** | **Arrival delay in minutes** |
| **16** |  | **DepDelay** | **Departure delay in minutes** |
| **17** |  | **Origin** | **Origin airport code** |
| **18** |  | **Dest** | **Destination airport code** |
| **19** |  | **Distance** | **Distance in miles** |
| **20** | **X** | **Taxiln** | **Taxi in time in minutes** |
| **21** |  | **TaxiOut** | **Taxi out in minutes** |
| **22** |  | **Cancelled** | **Was flight cancelled ?** |
| **23** |  | **Cancellationcode** | **Reason for cancellation**   1. **Carrier** 2. **Weather** 3. **Nas** 4. **Security** |
| **24** | **X** | **Diverted** | **1=yes 0=no** |
| **25** | **X** | **CarrierDelay** | **In minutes** |
| **26** | **X** | **WeatherDelay** | **In minutes** |
| **27** | **X** | **NASDelay** | **In minutes** |
| **28** | **X** | **SeurityDelay** | **In minutes** |
| **29** | **X** | **LateAircrafDelay** | **In minutes** |

The forbidden column in the table represents those variables from the dataset that can't be included since their values are only known once the plane has already taken off. Hence, if we use the forbidden variables, all the algorithms would return unreliable results.The forbidden columns include ArrTime, ActualElapsedTime, AirTime, TaxiIn,

Diverted, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay andLateAircraftDelay.

**See Below Dataset Schema**

root

|-- Year: integer (nullable = true)

|-- Month: integer (nullable = true)

|-- DayofMonth: integer (nullable = true)

|-- DayOfWeek: integer (nullable = true)

|-- DepTime: integer (nullable = true)

|-- CRSDepTime: integer (nullable = true)

|-- ArrTime: integer (nullable = true)

|-- CRSArrTime: integer (nullable = true)

|-- UniqueCarrier: string (nullable = true)

|-- FlightNum: integer (nullable = true)

|-- TailNum: string (nullable = true)

|-- ActualElapsedTime: integer (nullable = true)

|-- CRSElapsedTime: integer (nullable = true)

|-- AirTime: integer (nullable = true)

|-- ArrDelay: integer (nullable = true)

|-- DepDelay: integer (nullable = true)

|-- Origin: string (nullable = true)

|-- Dest: string (nullable = true)

|-- Distance: integer (nullable = true)

|-- TaxiIn: integer (nullable = true)

|-- TaxiOut: integer (nullable = true)

|-- Cancelled: integer (nullable = true)

|-- CancellationCode: string (nullable = true)

|-- Diverted: integer (nullable = true)

|-- CarrierDelay: integer (nullable = true)

|-- WeatherDelay: integer (nullable = true)

|-- NASDelay: integer (nullable = true)

|-- SecurityDelay: integer (nullable = true)

|-- LateAircraftDelay: integer (nullable = true)

**Methodology**

In this section, we review the dataset available to us and after proper cleaning of

the data we run different analytical methods to find insights of the data. For example, the various variables responsible for the arrival delay of flights.

After the exploratory analysis of the available data, we make the data run across certain predictive analysis methods like Logistic Regression and Random forest model and try to predict the flight arrival delay with the highest precision and accuracy .

Diagram

Description automatically generated

**Exploratory Data Analysis**

**Define Delay**

Our focus is to predict the value of ArrDelay ( arrival delay )

We define Arrival delay if the flight is late more than 15 minutes of its Scheduled Arrival

**Forbidden Variables**

In order to perform the required study properly, there are some

variables from the dataset that can't be included (forbidden variables) since their

values are only known once the plane has already taken off. If we used the forbidden

variables, all the algorithms would return unreliable results. For this reason, all the

columns ArrTime, ActualElapsedTime, AirTime, TaxiIn, Diverted, CarrierDelay,

WeatherDelay, NASDelay, SecurityDelay and LateAircraftDelay were dropped.

Reading the file

Graphical user interface, text, application, email

Description automatically generated

Background pattern

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**Correlation of features**

Chart

Description automatically generated

From the correlation Matrix we can see the strongest correlations are with DepDelay and TaxiOut , it also show some but much less correlation with depTime, CRSDEpTime and CRSArrTime.

**One Hot encoding Categorical Values**

A picture containing text

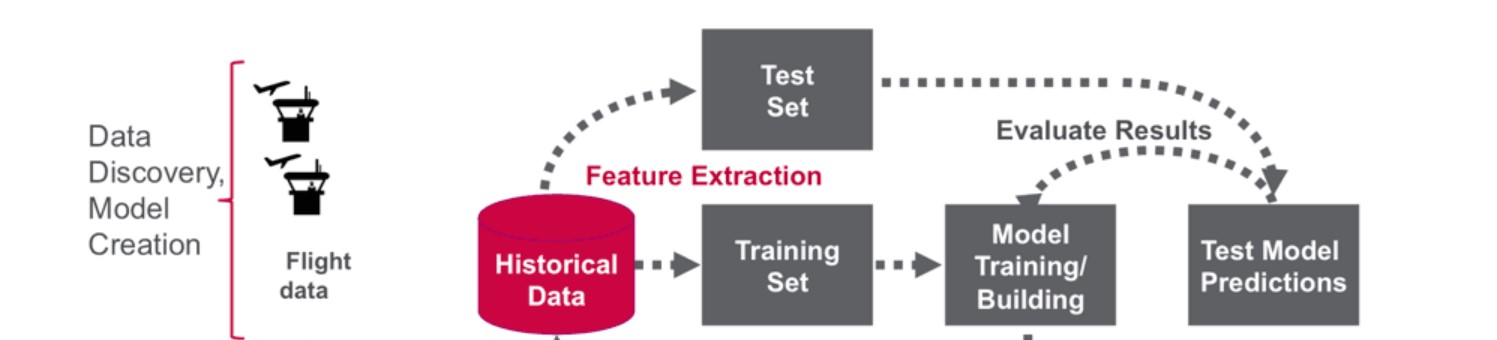
Description automatically generated

**Sampling and check for imbalance**

**Graphical user interface, text

Description automatically generated with medium confidence**

**Logistic Regression Model**

****

Text

Description automatically generated

Chart

Description automatically generated

**Evaluation of the Model.**

A picture containing shape

Description automatically generated

**Random Forest Model**

Graphical user interface, application, table

Description automatically generated

Conclusion.

Since we choose to run the logistic Regression Model using 12 features

And Also run the model with 5 features with the help of correlation matrix. The result are almost identical as at the end correlated features provide more accurate information for the model to predict.

When we run Random Forest model using 12 features we got a better result Auc , Recall and accuracy.

In general logistic regression performs better when the number of noise variables is less than or equal to the number of explanatory variables and random forest has a higher true and false positive rate as the number of explanatory variables increases in a dataset.

As Random forest model takes random samples forming many decision tress and take the average of those decisions to form a more refined model.

**References**

<https://stackoverflow.com/questions/58995226/one-hot-encode-of-multiple-string-categorical-features-using-spark-dataframes>

https://github.com/ssingh56/Predicting\_Flight\_delays\_cancellations/blob/master/EDA%2BDELAY.ipynb

<https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c>

<https://www.analyticsvidhya.com/blog/2019/11/build-machine-learning-pipelines-pyspark/>

<https://datagy.io/python-correlation-matrix/>