**PREDICTING FLIGHT DELAYS AND CANCELLATIONS**

**NADIPALLI SREE LEELA (820852403)**

**BIG DATA FINAL PROJECT**

**INTRODUCTION**

This project is on detecting the flight cancellation, arrival and departure delays of ***San Diego Airport.*** Based on different parameters in the dataset, the cancellation and delays are determined.

**DATASET**

The dataset flights.csv is taken from kaggle and the link to the dataset is given here. <https://www.kaggle.com/usdot/flight-delays/data>

The data set has 31 rows initially. However two rows are added with the help of excel formulae to increase the speed and accuracy. Final dataset has 33 rows.

ARRIVAL\_DELAY\_INDEX and DEPARTURE\_DELAY\_INDEX columns are added. If flight is delayed, then value 0 is given else 1 is given. Based on timings given in departure delay and arrival delay columns, it is determined if flight is delayed. 0 is given if flight is cancelled.

The table below shows the final dataset:

|  |  |  |
| --- | --- | --- |
| **S.NO** | **DATA FIELD** | **DATATYPE** |
| 1 | YEAR | INTEGER |
| 2 | MONTH | INTEGER |
| 3 | DAY | INTEGER |
| 4 | DAY\_OF\_THE\_WEEK | INTEGER |
| 5 | AIRLINE | STRING |
| 6 | FLIGHT\_NUMBER | STRING |
| 7 | TAIL\_NUMBER | INTEGER |
| 8 | ORIGN\_AIRPORT | STRING |
| 9 | DESTINATION\_AIRPORT | STRING |
| 10 | SCHEDULED\_DEPARTURE | INTEGER |
| 11 | DEPARTURE\_TIME | INTEGER |
| 12 | DEPARTURE\_DELAY | INTEGER |
| 13 | DEPARTURE\_DELAY\_INDEX | INTEGER |
| 14 | TAXI\_OUT | INTEGER |
| 15 | WHEELS\_OFF | INTEGER |
| 16 | SCHEDULED\_TIME | INTEGER |
| 17 | ELAPSED\_TIME | INTEGER |
| 18 | AIR\_TIME | INTEGER |
| 19 | DISTANCE | INTEGER |
| 20 | WHEELS\_ON | INTEGER |
| 21 | TAXI\_IN | INTEGER |
| 22 | SCHEDULED\_ARRIVAL | INTEGER |
| 23 | ARRIVAL\_TIME | INTEGER |
| 24 | ARRIVAL\_DELAY | INTEGER |
| 25 | ARRIVAL\_DELAY\_INDEX | INTEGER |
| 26 | DIVERTED | INTEGER |
| 27 | CANCELLED | INTEGER |
| 28 | CANCELLED\_REASON | INTEGER |
| 29 | AIRSYSTEM\_DELAY | INTEGER |
| 30 | SECURITY\_DELAY | INTEGER |
| 31 | AIRLINE\_DELAY | INTEGER |
| 32 | LATE\_AIRCRAFT\_DELAY | INTEGER |
| 33 | WEATHER\_DELAY | INTEGER |

For better performance data is taken is taken for one month. However, data for five months is tested and results are discussed in observations section.

***The dataset from kaggale contains all airports data. However, in this project only San Diego airport data is studied.***

Total number of rows for one month are **12170**.

In this project Random Forest Model and Logistic Regression are implemented. And the results are compared.

Observations made by other people are also compared and studied.

**PREDICTIONS**

In each algorithm three observations are predicted:

* Arrival Delays
* Departure Delays
* Cancellations

The output is ran ten times and average results are discussed for accuracy, to get better sense of algorithms.

**RANDOM FOREST MODEL**

* Cancellations

Confusion Matrix

+------+------+

|y\_is\_0|y\_is\_1|

+------+------+

| 0.0| 0.0|

| 36.0|3632.0|

+------+------+

Random Forest Cancellation Accuracy: 99.0185387131952

* Departure Delays

Confusion Matrix

+------+------+

|y\_is\_0|y\_is\_1|

+------+------+

| 88.0| 27.0|

|1334.0|2117.0|

+------+------+

Random Forest Departure delay accuracy: 61.8339876612451

* Arrival Delays

Confusion Matrix

+------+------+

|y\_is\_0|y\_is\_1|

+------+------+

| 0.0| 0.0|

|1364.0|2320.0|

+------+------+

Random Forest Arrival delay accuracy: 62.975027144408244

**LOGISTIC REGRESSION MODEL**

* Cancellations

Confusion Matrix

+------+------+

|y\_is\_0|y\_is\_1|

+------+------+

| 0.0| 0.0|

| 28.0|3571.0|

+------+------+

LR Cancellation Accuracy: 99.22200611280911

* Departure Delays

Confusion Matrix

+------+------+

|y\_is\_0|y\_is\_1|

+------+------+

| 570.0| 411.0|

| 854.0|1851.0|

+------+------+

LR Departure Delays Accuracy: 65.68095496473141

* Arrival Delays

Confusion Matrix

+------+------+

|y\_is\_0|y\_is\_1|

+------+------+

| 146.0| 135.0|

|1182.0|2223.0|

+------+------+

LR Arrival Delays Accuracy: 64.27021161150299

**OBSERVATIONS**

From the above observations it is clear that Logistic Regression has better accuracy when compared to Random Forest Model. However, there is no drastic difference between the two.

This is because we have taken limited dataset. If we consider three months data, then the accuracy increased 5% more and if we consider five months data then accuracy increased 10% more in Logistic Regression. And even considerable amount of accuracy difference between the two models can been seen.

On smaller dataset both algorithms take almost same time. However, if data increases, logistic regression is little faster.

However, if we consider entire 2.5 million datasets then the accuracy reaches over 80%. But, it is to be noted that as accuracy increases time and space complexity increases as we have to consider large dataset.

There are several experiments performed on this dataset by various other people, and the results obtained here and the results discussed in kaggale forums are nearly equal.

**OUTPUT**

We get the following file structure as output

* LogisticRegressionModel
  + ArrivalDelays
    - confusionMatrix
    - model
  + DepartureDelays
    - confusionMatrix
    - model
  + Cancellations
    - confusionMatrix
    - model
* RandomForestModel
  + ArrivalDelays
    - confusionMatrix
    - model
  + DepartureDelays
    - confusionMatrix
    - model
  + Cancellations
    - confusionMatrix
    - model

**PROJECT DEVELOPMENT ENVIRONMENT**

The project is done in IntelliJ.

Spark 2.11.11 is used

SBT is used to get jar file.