SPE International

University of Houston
SPE Student Chapter

Machine Learning Bootcamp - 2020

Team No. 4

Project 2 Final Report

Submitted By-
Fatmir Likrama- Business Analyst
Pratik Ghatake- Developer
Miguel Mendoza – Data Engineer
Celine Cherian – Data Scientist

1. Problem Introduction

As airline travel has become economical for most of the world's population it has also become increasingly popular with the number of passengers showing year by year increase even in economic downturns. It has become irreplaceable but despite all the conveniences it provides, it can be agreed that we spend more time than we'd want in airports, whether waiting four our loved ones' arrival or catching a flight to another destination. One (probably the most important) reason for this waiting and time lost is flight delays.

In this project we use data analysis, statistics and machine learning techniques to gain insights into the most important parameters and ultimately to predict whether a flight will be delayed or not based on several relevant factors.

2. Dataset

The dataset used for this project is flight data for the month of January 2019, a large dataset with more than 58k rows of non-null data. The data was mostly categorical in nature and consisted of columns for day of week, day of month, carrier ID, origin and destination ID, flight tail number, departure and arrival time blocks, whether a flight was diverted or canceled, distance traveled and whether it was delayed or not. Our target variable is whether a flight was delayed or not.

3. Features and Processing

Some processing was applied to the data mainly to drop features not considered relevant to predicting the target variable such as flight tail number, flight number, arrival time and duplicate information such as destination and origin codes and carrier codes. Rows with null data were also excluded

Exploratory plots of data showed that whether a flight is delayed or not would depend on day of week, day of month, carrier ID but also on distance traveled, departure time block, origin and destination airport and whether a flight was late in departure or not. These features were ultimately used to train models and predict whether a flight would arrive late or not. The size of the data resulted in high computation times.

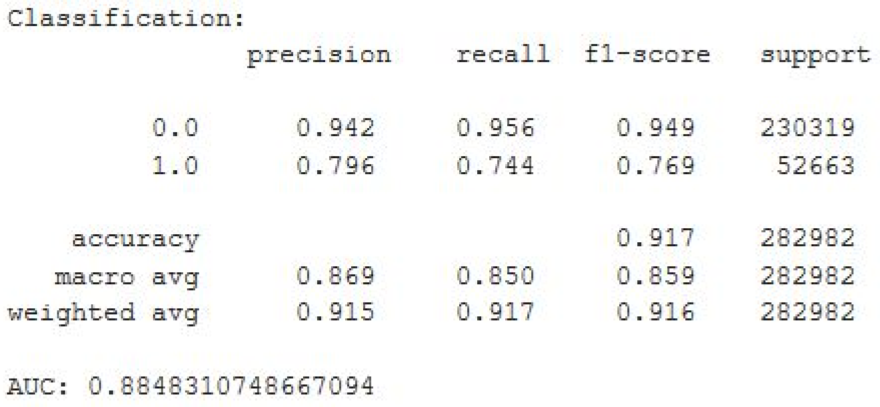
4. Models and Techniques

The following machine learning techniques were used:

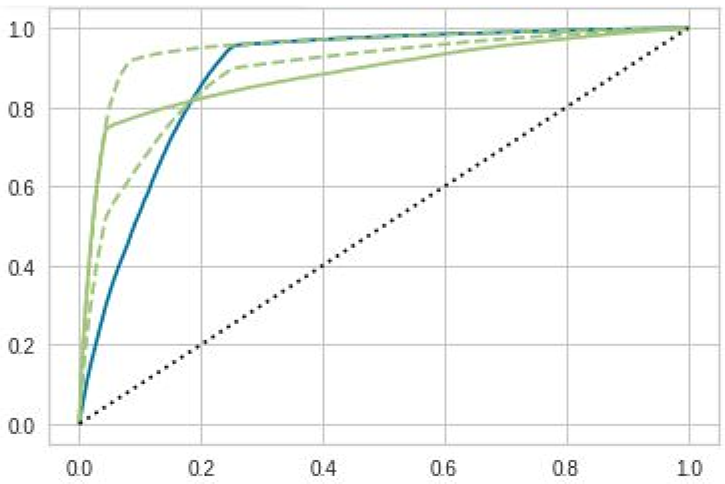
- **Logistic Regression**
- **Decision Tree Classifier**
- **Random Forest Classifier**

5. Results

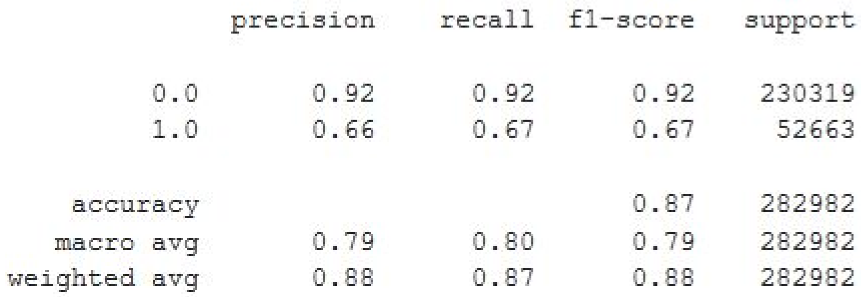
- **Logistic Regression**



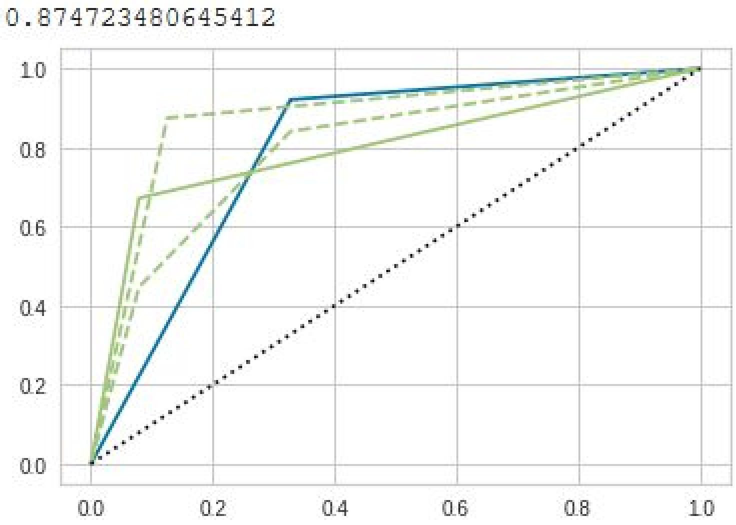
```
Classification:
              precision    recall  f1-score   support

         0.0      0.942     0.956     0.949    230319
         1.0      0.796     0.744     0.769     52663

    accuracy                          0.917    282982
   macro avg      0.869     0.850     0.859    282982
weighted avg      0.915     0.917     0.916    282982

AUC: 0.8848310748667094
```



- **Decision Tree Classifier**



```
              precision    recall  f1-score   support

         0.0       0.92      0.92      0.92    230319
         1.0       0.66      0.67      0.67     52663

    accuracy                           0.87    282982
   macro avg       0.79      0.80      0.79    282982
weighted avg       0.88      0.87      0.88    282982
```



```
0.874723480645412
```

Summary and Conclusions

We used data analytics, statistics and machine learning methods to predict whether a flight would be late or not. Logistic Regression, Decision Tree, and Random Forest classifier methods were used with satisfactory results. Logistic Regression performed better than Decision Tree Classifier method but the AOC of ROC curve for the two methods are very similar.

The non perfect match between model prediction and reality as seen in test data validation indicates that additional parameters affect whether a flight is delayed or not. For example, weather and airport logistics would definitely be important factors. Including these extra factors would improve our prediction.

References

Data: https://www.kaggle.com/divyansh22/flight-delay-prediction