BY VENKATA SUMANTH TEJA

SUBMITTED TO: PROF. ANDREW ENKEBOLL

PREDICTING FLIGHT DELAYS

Problem statement

- Flight delay is inevitable, and it plays an important role in both profits and loss of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and incomes of airline agencies. There have been many researches on modeling and predicting flight delays, where most of them have been trying to predict the delay through extracting important characteristics and most related features. However, most of the proposed methods are not accurate enough because of massive volume data, dependencies and extreme number of parameters.

DATA EXPLORATION

- Dataset : https://www.kaggle.com/code/asrsaiteja/introduction-to-pyspark-using-flightsdata/data
- I had taken the dataset from the Kaggle and dataset consists of 275000 rows and 10 columns.
- Data types are:

```
mon           int64
dom           int64
dow           int64
carrier      object
flight        int64
org          object
mile          int64
depart      float64
duration      int64
delay       float64
dtype: object
```

Data distribution of the dataset

	mon	dom	dow	flight	mile	depart	duration	delay
count	275000.000000	275000.000000	275000.000000	275000.000000	275000.000000	275000.000000	275000.000000	258289.000000
mean	5.242320	15.714069	2.946091	2063.054276	881.222287	14.124931	151.641036	28.347731
std	3.427357	8.805568	1.963514	2185.852170	700.517889	4.683190	87.084564	54.014895
min	0.000000	1.000000	0.000000	1.000000	11.000000	0.120000	14.000000	-80.000000
25%	2.000000	8.000000	1.000000	425.000000	342.000000	10.000000	85.000000	-6.000000
50%	5.000000	16.000000	3.000000	1079.000000	651.000000	14.080000	125.000000	15.000000
75%	8.000000	23.000000	5.000000	2785.000000	1162.000000	18.080000	192.000000	43.000000
max	11.000000	31.000000	6.000000	6941.000000	4243.000000	23.980000	605.000000	1370.000000

Problem solution

- I shall be going to use the exploratory data analysis(EDA) in this project and also going to do prediction of the flights in the given year.


THANK YOU