Predicting Flight Delay Using Machine Learning

Presentation outline

- Introduction and objectives
- Data gathering & Pre-processing
- Exploratory Data Analysis
- ML – Modelling
- Summary

Introduction & Objectives

- Flight delays has become a very important subject for air transportation all over the world because of the associated financial loses that the aviation industry is continuously going through.
- According to the Bureau of Transportation Statistics (BTS) of the US, over 20% of the US flights were delayed during 2018 41 billion US$.
- Delays caused inconvenience to airlines and passengers Financial loses and increased stress.
- Is it possible to predict when a flight will be delayed even before it comes out in the departure board?

Objective: Design a model that predicts flight delays before they are announced on the departure boards.

Data Gathering & Pre-processing

Full Dataset:

- ➢Source: Kaggle
- ➢Data from 10 years (2009 – 2018).
- ➢10 different files.
- ➢Average of 28 Categories (> 1 million rows) .

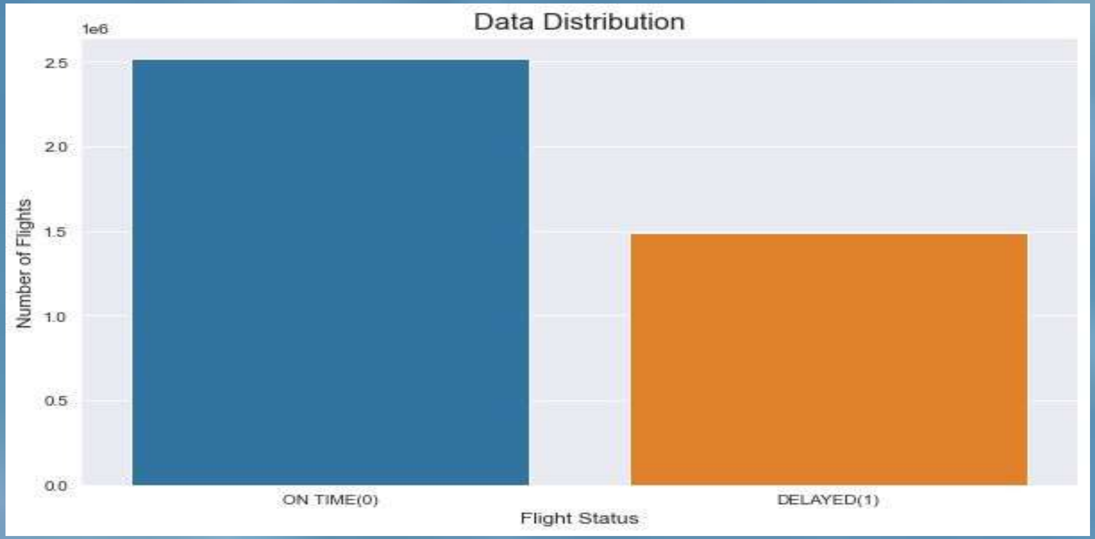
Selected Dataset:

- ➢1 year: 2018
- ➢ +7.2 million rows : 20 top destinations.

(cities): +4.2 million rows

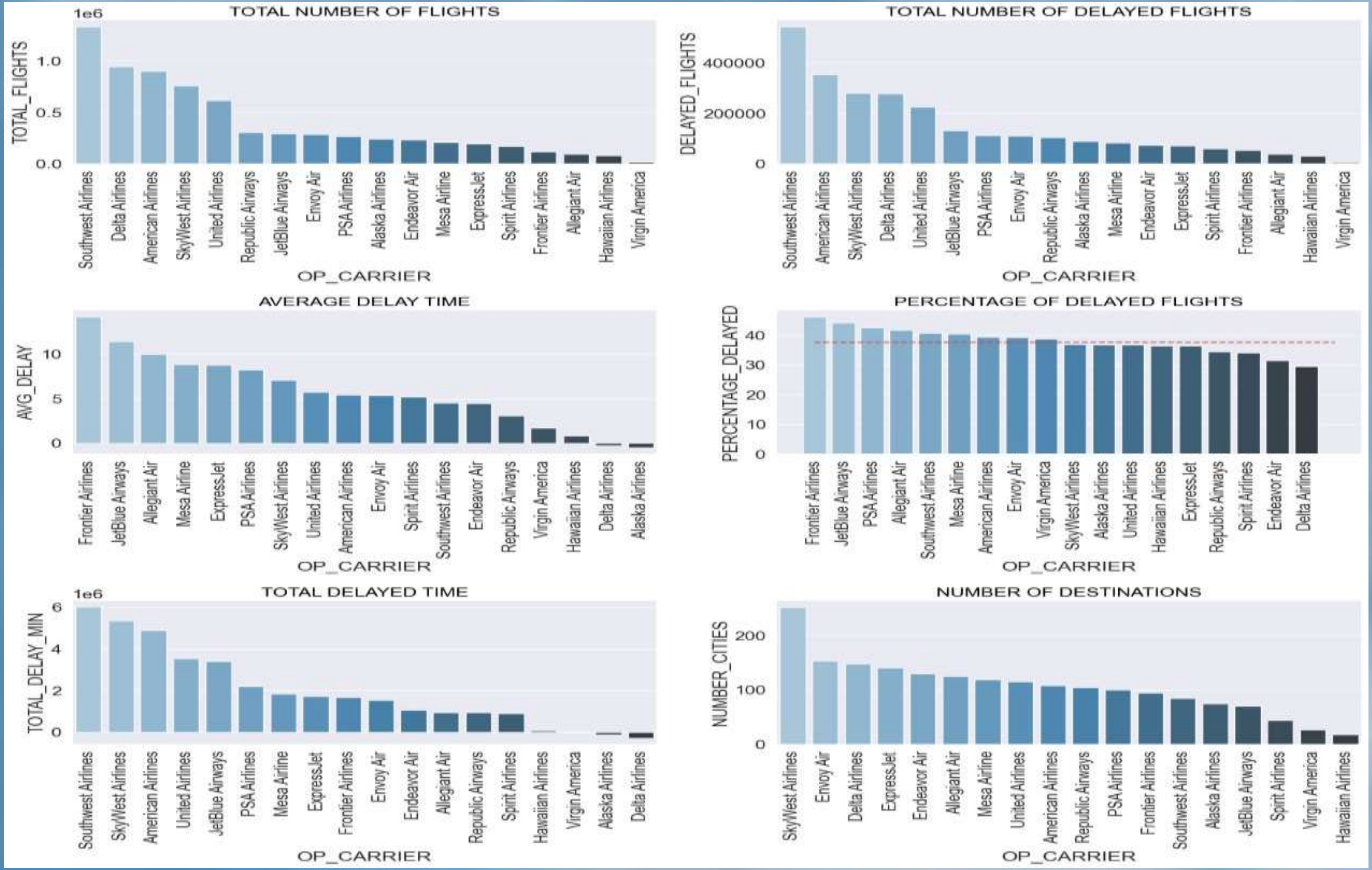
What differentiates my models from others:

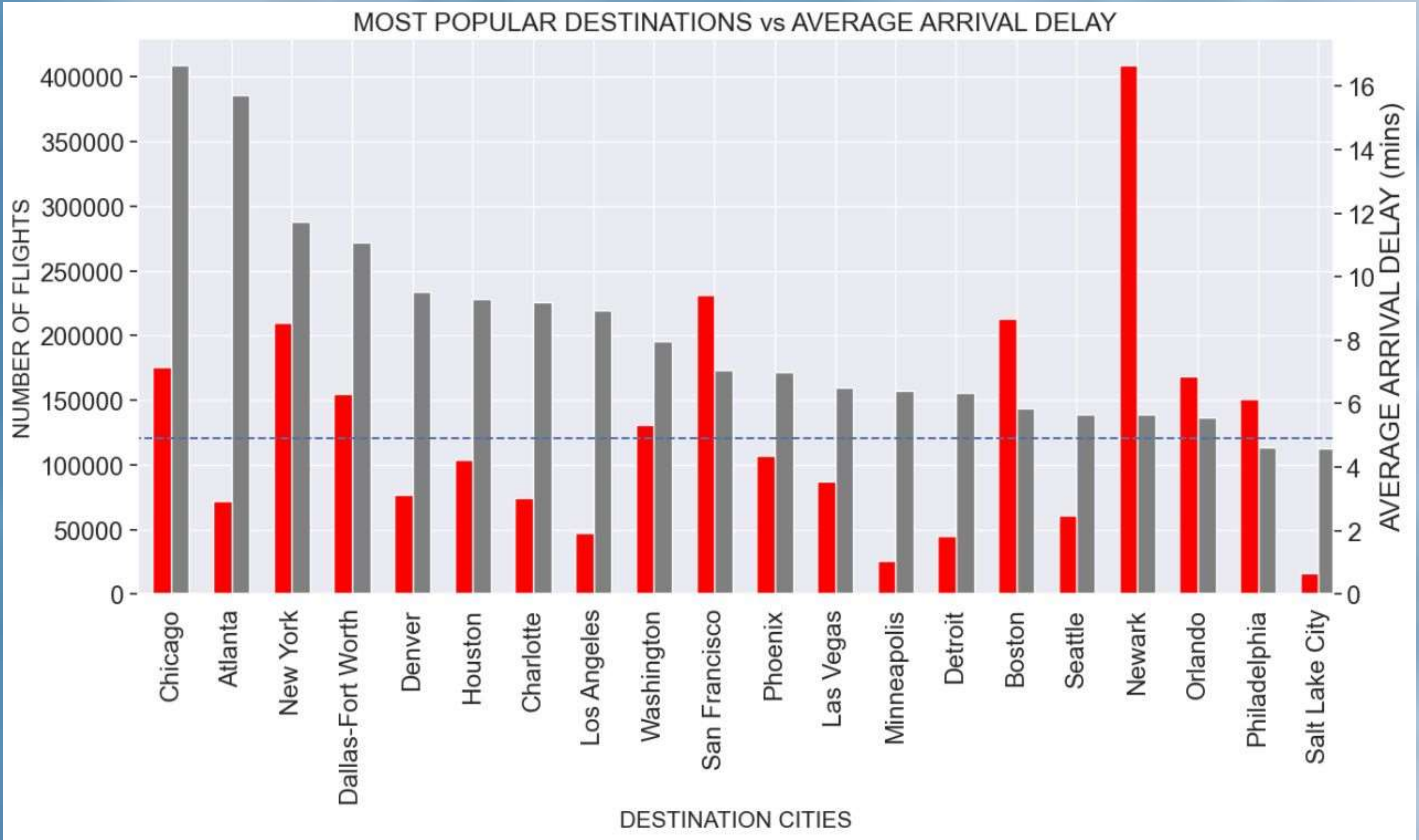
- No category that would imply passengers already in the plane or the delay announced on departure boards were considered.
- Many Categorical (days, months, origin, destination, times,…)
- Imbalance with almost a 2:1 ratio:



- Over 15 different features were engineered for Exploratory Data Analysis

Exploratory Data Analysis




MOST POPULAR DESTINATIONS vs AVERAGE ARRIVAL DELAY
NUMBER OF FLIGHTS
400000
350000
300000
250000
200000
150000
100000
50000
0
AVERAGE ARRIVAL DELAY (mins)
16
14
12
10
8
6
4
2
0
Chicago
Atlanta
New York
Dallas-Fort Worth
Denver
Houston
Charlotte
Los Angeles
Washington
San Francisco
Phoenix
Las Vegas
Minneapolis
Detroit
Boston
Seattle
Newark
Orlando
Philadelphia
Salt Lake City
DESTINATION CITIES

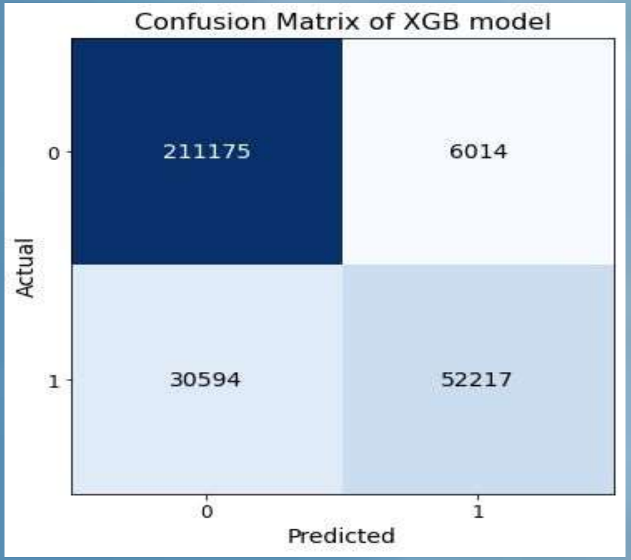
ML – Modelling

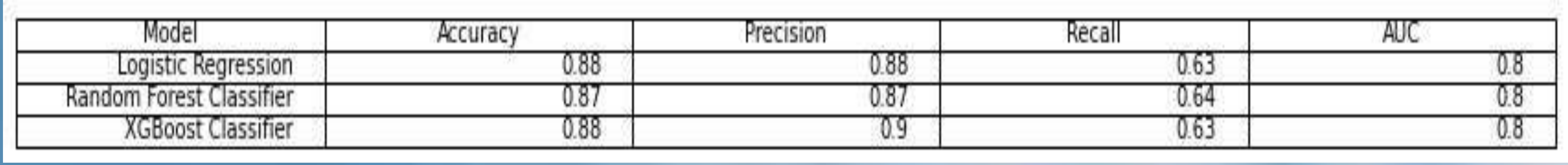
Binary Classification:

- 0 = Flight arrives on-time
- 1 = Delayed Flight

Algorithms tested:

- Logistic Regression
- Random Forest
- XGBoost





Model	Accuracy	Precision	Recall	AUC
Logistic Regression	0.88	0.88	0.63	0.8
Random Forest Classifier	0.87	0.87	0.64	0.8
XGBoost Classifier	0.88	0.9	0.63	0.8

Summary

- From the EDA done it seems like Delta Airlines and Alaska Airlines are two of the most reliable airlines in terms of flights arriving on time.
- It is quite hard to create a ML model to predict flight delays without giving them any features that could affect the models by biasing them.
- The best model ended up with an accuracy of over 88%, however a series of categories believed to be key could not be considered due to a shortage of data. Adding these could increase the accuracy and other metrics.

Thank you....