Analysis of Delayed Flights

Janith Perera

Introduction

- Aviation industry plays a critical role in the World's transportation sector.
- Lot of business relay various airlines to connect them with other parts of the world.
- But several reasons may directly affect the airline services by means of flight delays.
- Accurately predicting these delays will be beneficial to passengers to be well prepared and to airlines to respond to the potential causes of the flight delays in advance.

Purpose of the project

- Build a model for predicting flight delays at the destination airport specifically for the month of January in upcoming years.
- Exploratory data analysis on the variables to find out the tends and role in determining the flight delays.
- Compare the prediction accuracy of the target variable which is arrival delays with using popular classification algorithms which are Logistics Regression, Stochastic Gradient Descent, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting and XB Boost.

Problem Statement

Using popular classification techniques, can we predict the possible flight delays based on factors like day of the week, date of the month, carrier, origin and destination etc.

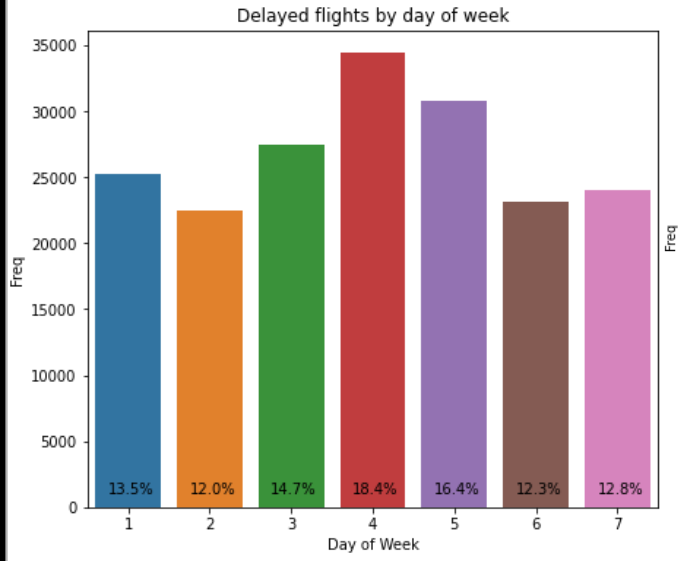
Data

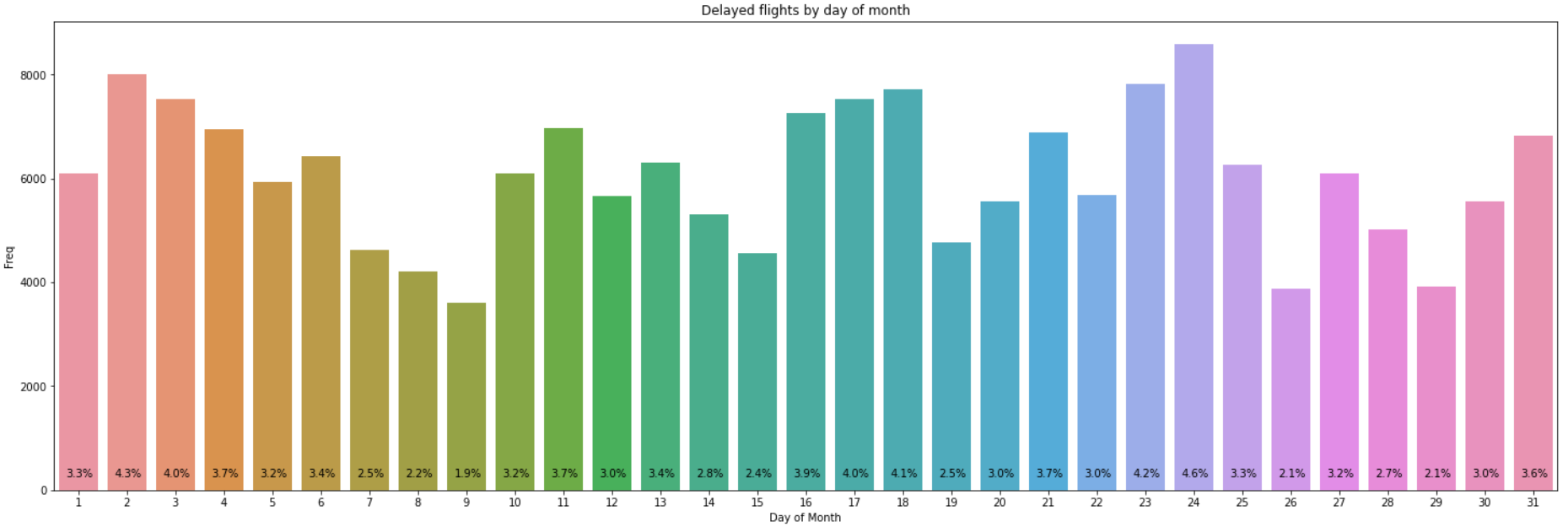
- Downloaded from Kaggle.com
- Originally collected from the Bureau of Transportation Statistics, Govt. of the USA.
- Data contains all the flights throughout the United States in January 2019 and January 2020.
- Includes 21 feature columns such as origin airport, destination airport, airplane information, departure time, arrival time and etc.
- About 1.2 million flight information (instances)
- Binary target variable.

Exploratory Data Analysis

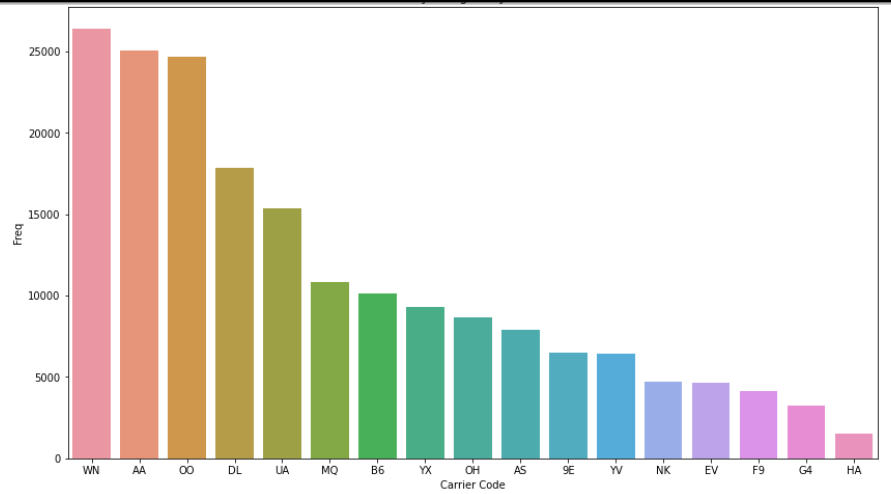
DEST	ARR_DEL15	PERCENT
ORD	10170.0	5.423798
DFW	8667.0	4.622227
ATL	7263.0	3.873455
LGA	7077.0	3.774259
SFO	6114.0	3.260678

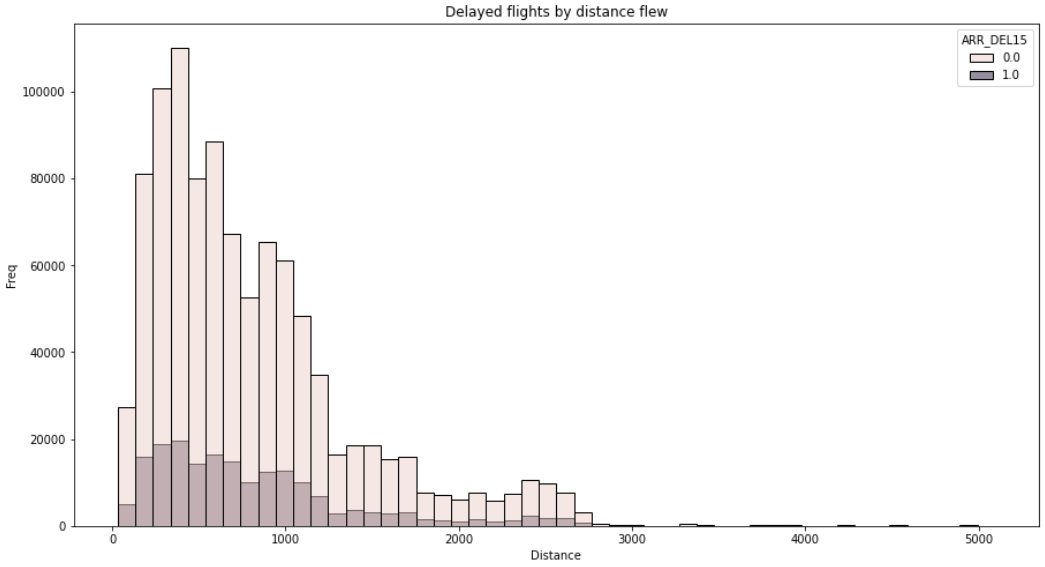
ORIGIN	DEP_DEL15	PERCENT
ORD	10736.0	5.937692
DFW	8597.0	4.754689
ATL	7784.0	4.305048
DEN	6195.0	3.426230
CLT	5744.0	3.176798

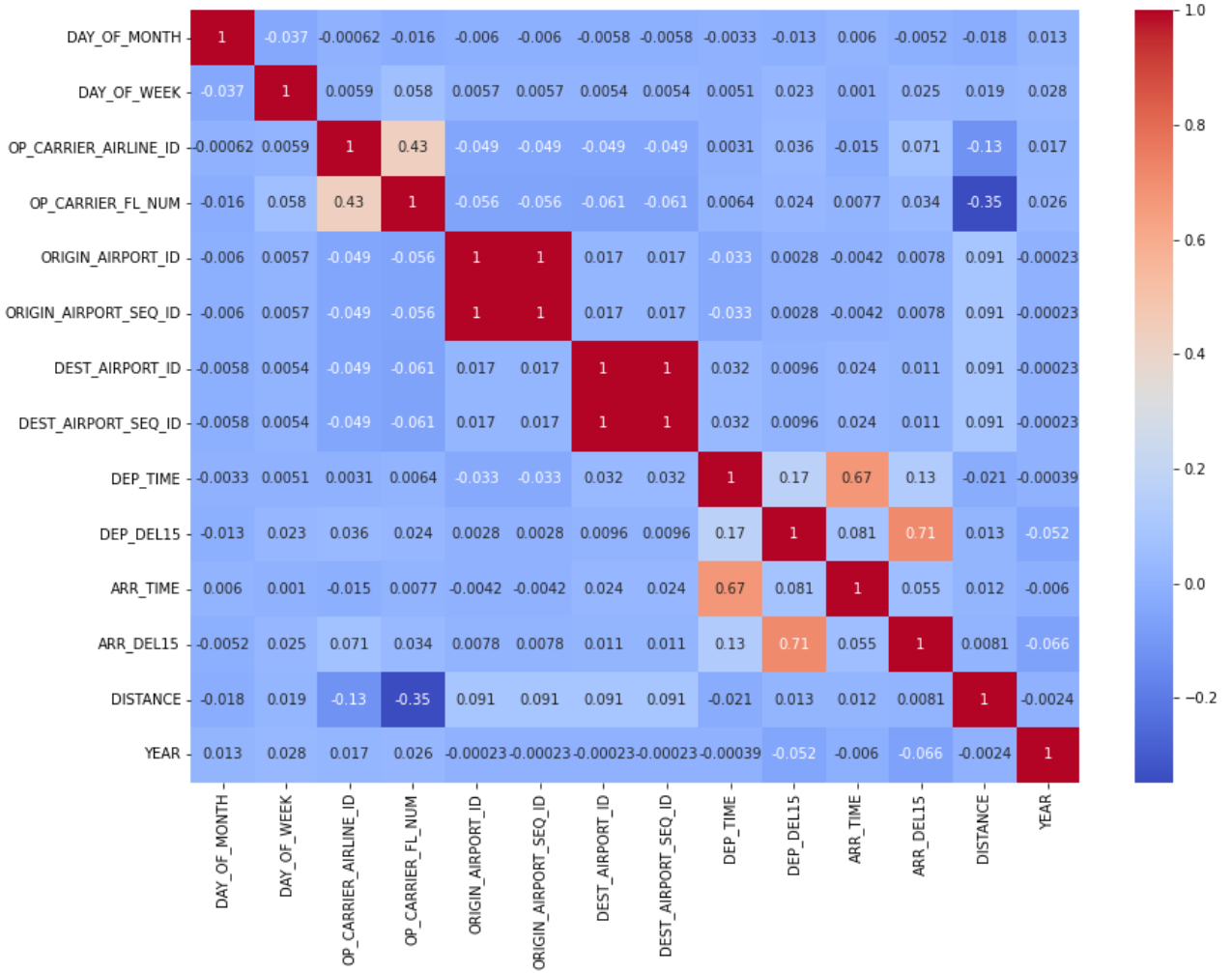




Exploratory Data Analysis







Dataset Preparation

- Cancelled and diverted flights considered as delayed since those flights are negligible when considered to total dataset.
- Rest of null values were less than 3% and dataset quit large, therefore flights with missing data removed.

TAIL_NUM	0.27%
DEP_TIME	1.97%
DEP_DEL	1.97%
ARR_TIME	2.07%

- Departure time and arrive time features transformed in to the time blocks.
- Created separate features for origin and destination airport sequence numbers.

Dataset Preparation

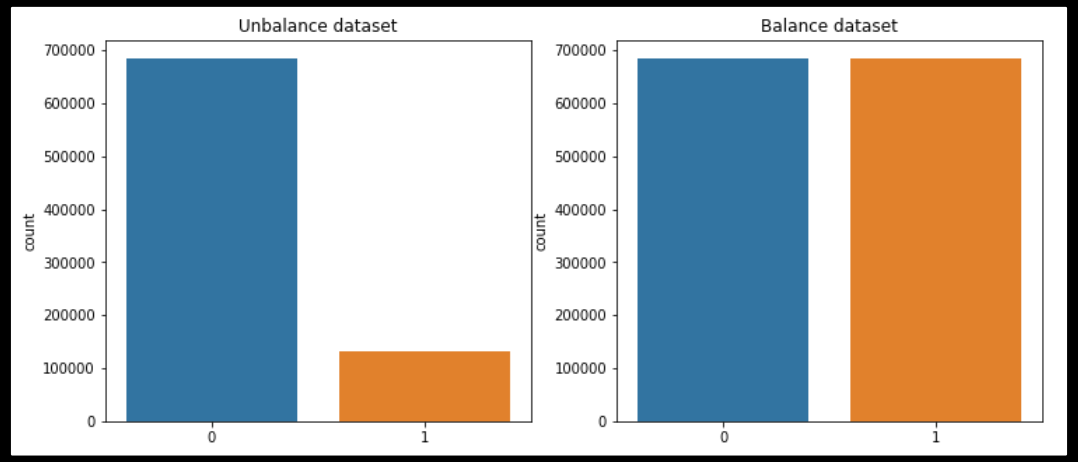
- Some features dropped to reduce correlation between features

Dropped features	Reason
Unnamed: 21	Do not carry any data
CANCELLED, DIVERTED	Data merged to target feature
DEP_TIME, ARR_TIME	Transformed in to time blocks
OP_CARRIER	Same data as OP_UNIQUE_CARRIER
OP_CARRIER_AIRLINE_ID	OP_UNIQUE_CARRIER is a subset of this
ORIGIN_AIRPORT_ID, ORIGIN_AIRPORT_SEQ_ID, DEST_AIRPORT_ID, DEST_AIRPORT_SEQ_ID	Created separate features

- Changed categorical data into categorical data type

Dataset Preparation

- LabelEncoder() was applied for all categorical features.
- LabelBinarizer() was applied to the target feature.
- RandomOverSampler() was applied to balance the dataset.



Model Performance

- Due to whole dataset is quite large, it took unreasonable time in local computer. Therefore, random 10% faction used to evaluate the models.
- For initial evaluation, 5-fold cross validation model used with seven popular classification models.

	Logistics Regression	SGD Classifier	Decision Tree Classifier	Random Forest Classifier ★	SVC	Gradient Boosting Classifier	XG Boost Classifier
Accuracy	0.85	0.62	0.95	0.98	0.53	0.85	0.90
Precision	0.95	0.72	0.91	0.96	0.53	0.95	0.93
Recall	0.74	0.61	0.99	0.99	0.53	0.74	0.86
F1 score	0.83	0.56	0.95	0.98	0.53	0.83	0.90

- Among those models, Random Forest classification model selected to hyperparameter tuning.

Results

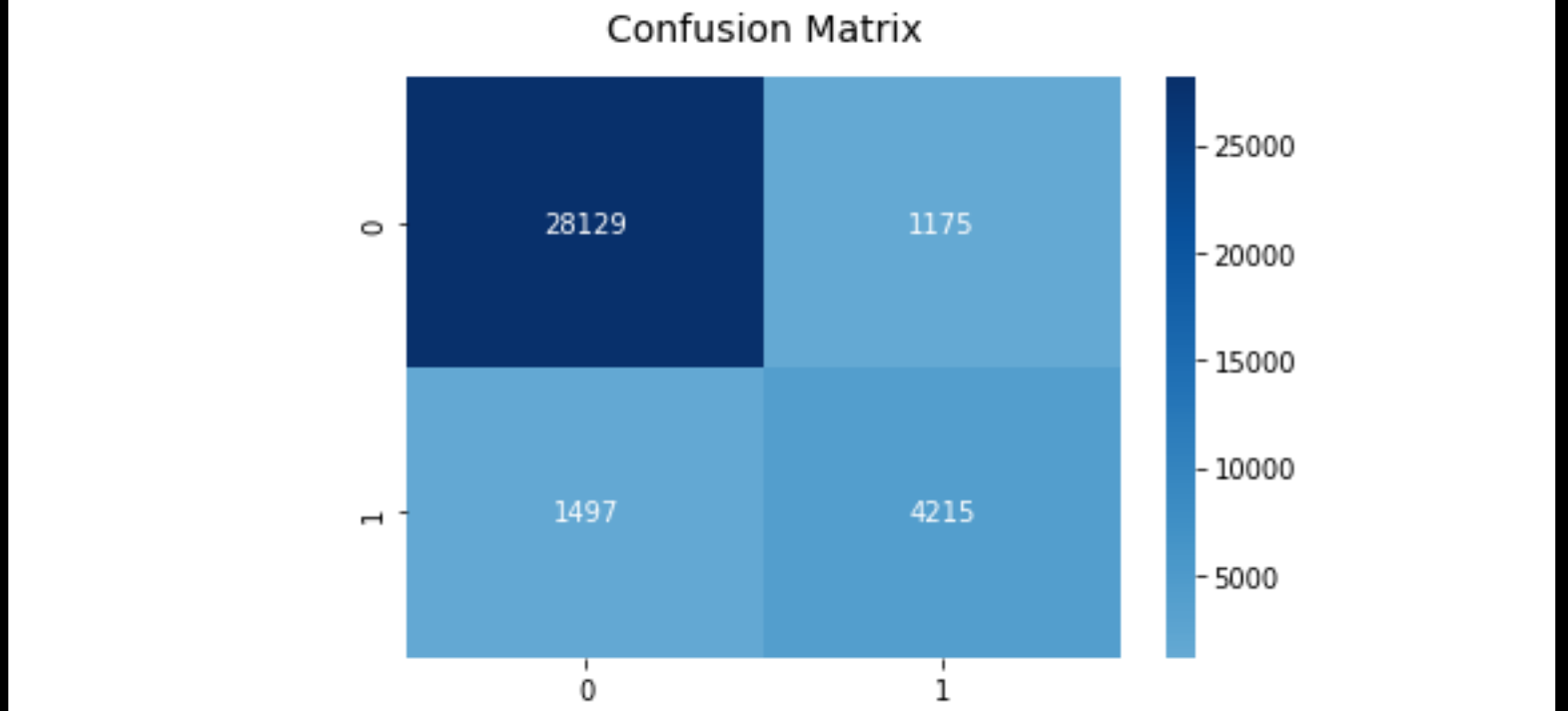
- Hyperparameter tuned Random Forest classification model gives the overall accuracy score of 1.00 on training set and 0.92 on testing set.

Classification report on training set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	68310
1	1.00	1.00	1.00	68310
accuracy			1.00	136620
macro avg	1.00	1.00	1.00	136620
weighted avg	1.00	1.00	1.00	136620

Classification report on testing set

	precision	recall	f1-score	support
0	0.95	0.96	0.95	29304
1	0.78	0.74	0.76	5712
accuracy			0.92	35016
macro avg	0.87	0.85	0.86	35016
weighted avg	0.92	0.92	0.92	35016



Conclusion and Lessons Learned

- I would recommend using a hyperparameter tuned Random Forest Classifier to assist in predicting flight delays at the destination airport for the month of January in upcoming years based on physical factors.
- It is better to include weather conditions in to this analysis because weather is one of the most influencing factor for flight delays.
- Further, if we can consider flights flew in whole year, we may find some patterns of delaying.

Thank you