Flight Delay Prediction

Nicole Hong
Siyu Mao

March 18, 2022

Project Importance

- 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 going through.
- According to data from the Bureau of Transportation Statistics (BTS) of the United States, over 20% of US flights were delayed during 2018, which resulted in a severe economic impact equivalent to 41 billion USD.
- The result is an increase in travel time which increases the expenses associated with food and lodging and ultimately causes stress among passengers.

Reference:
https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c

Project Overview

01 Workflow

02 EDA Insights

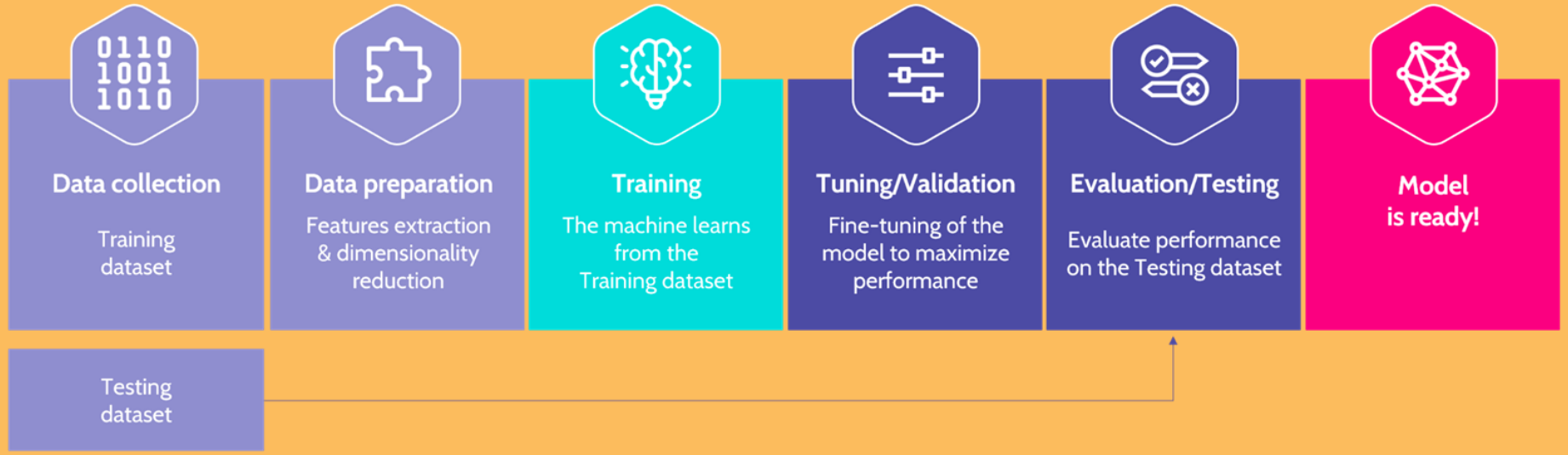
03 Data Preparation & Feature Engineering

04 Machine Learning Model & Performance

01

Project Workflow

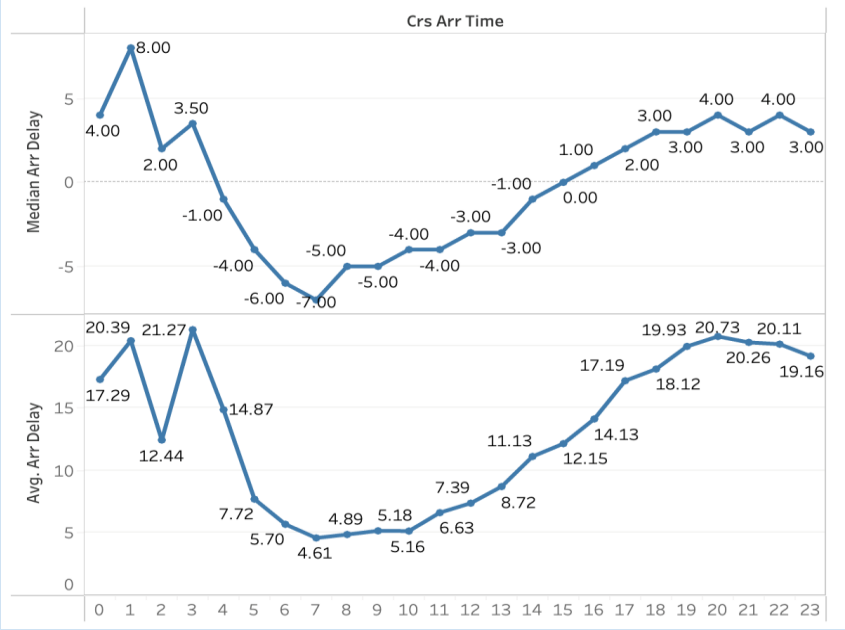
Workflow Chart

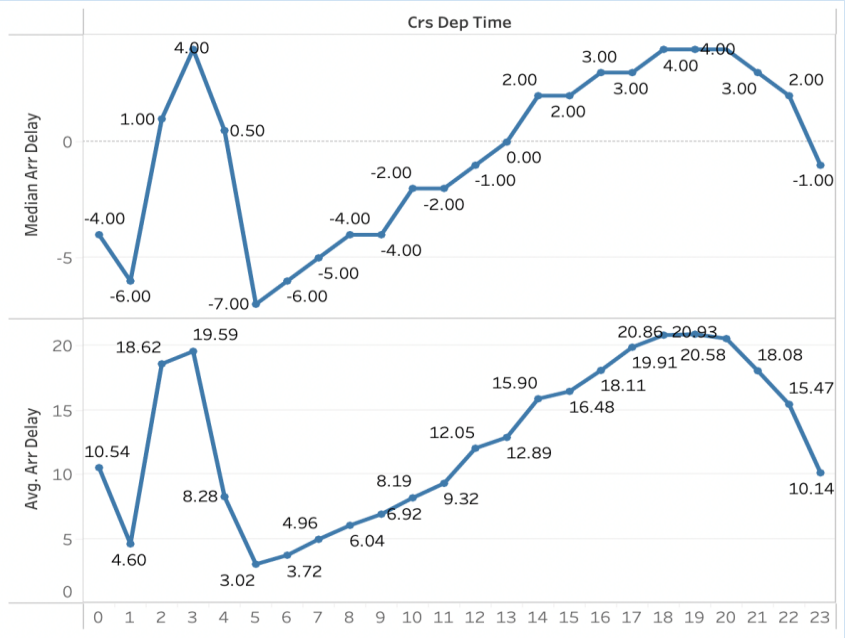


02

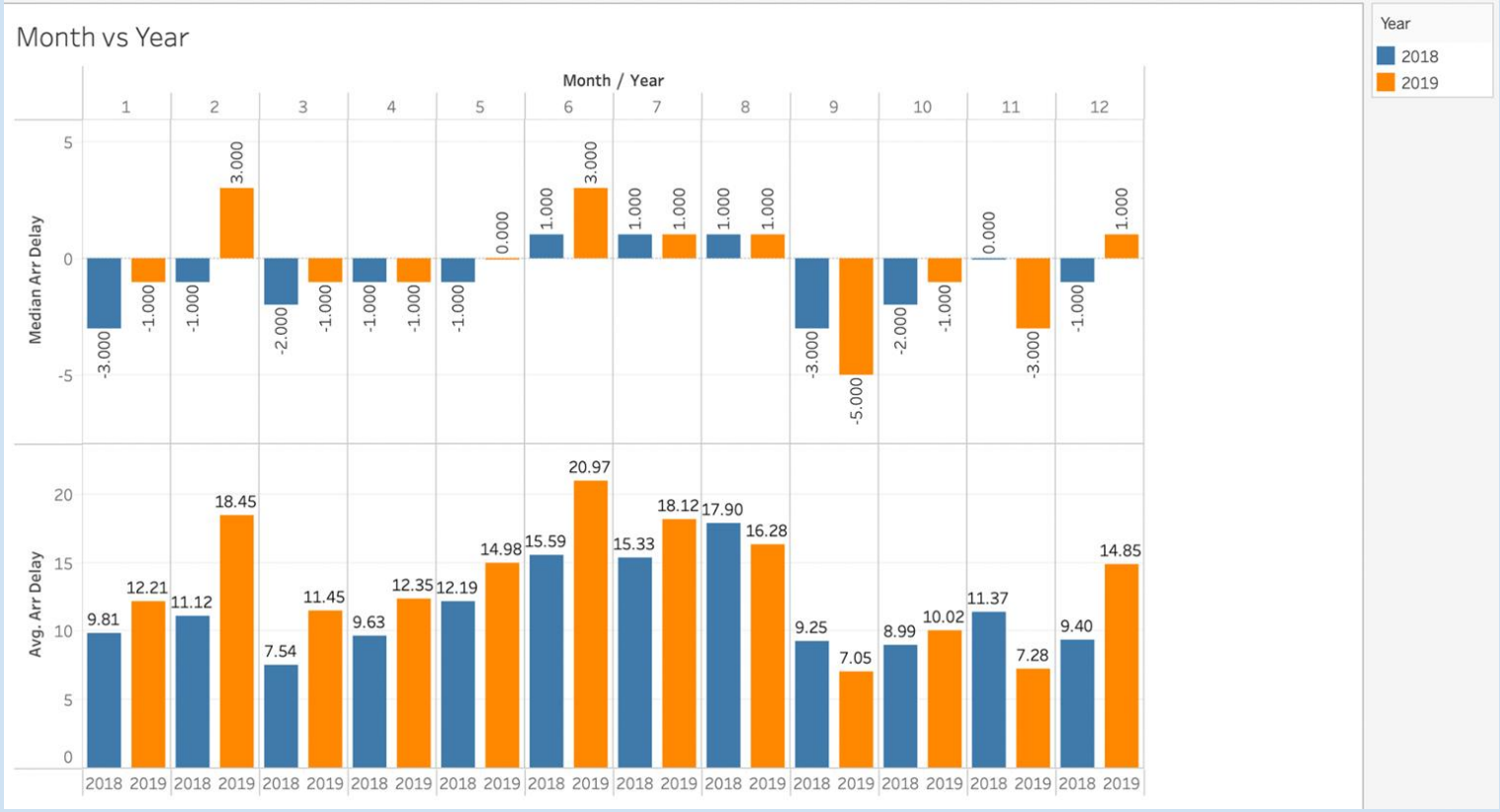
EDA Insights

Flights (fl_date) - Hour

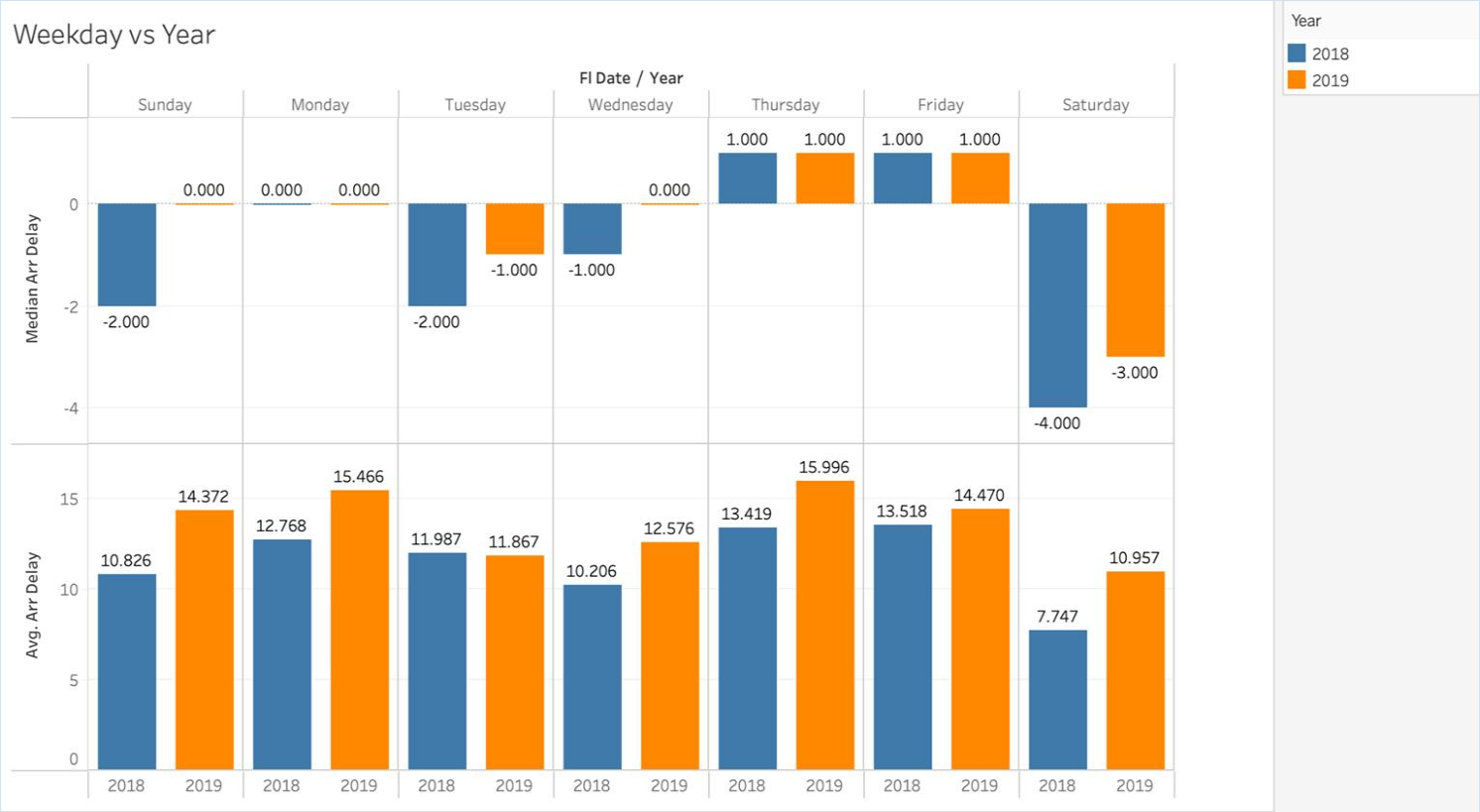




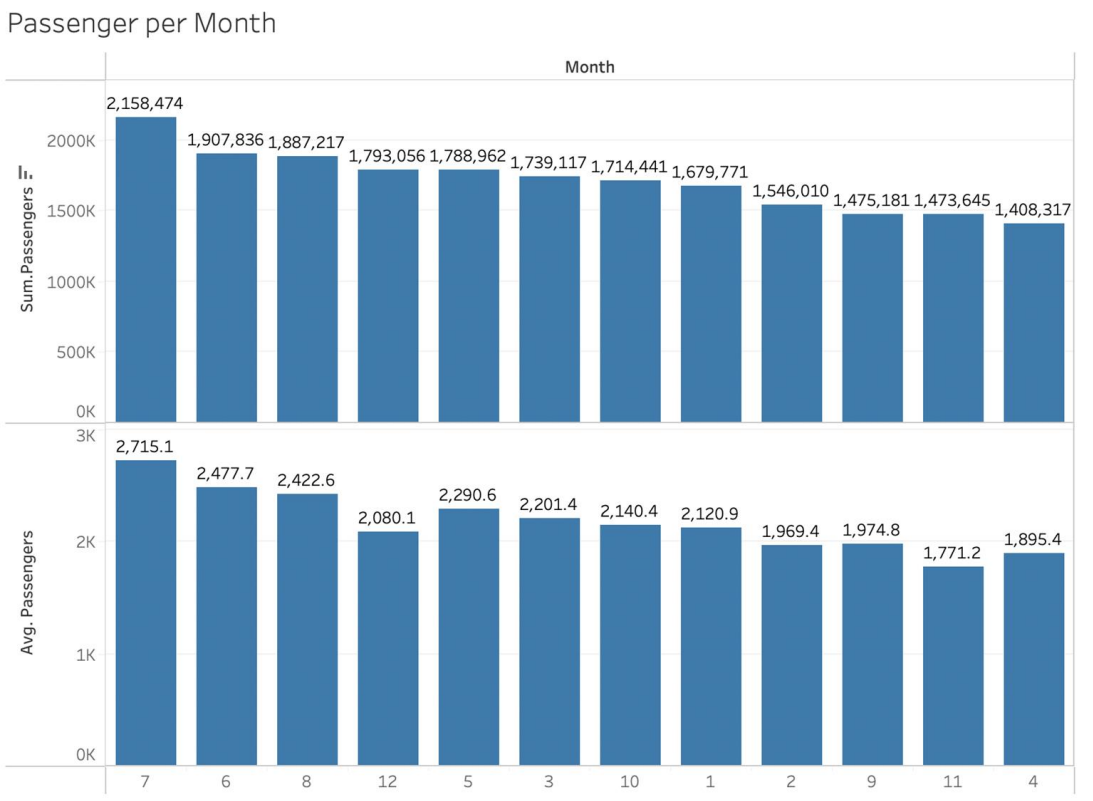
Flights (fl_date) - Month



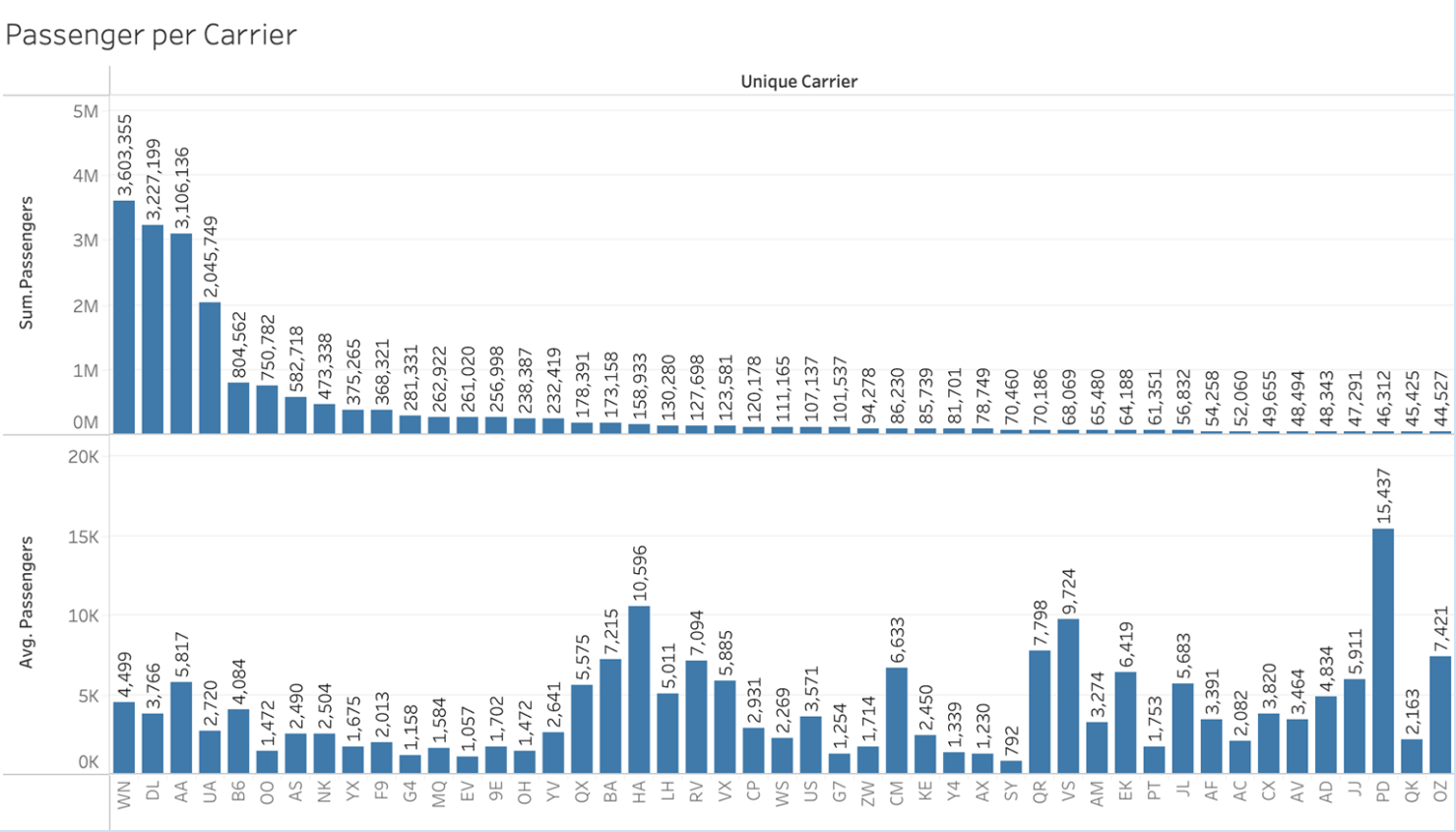
Flights (fl_date) - Weekday



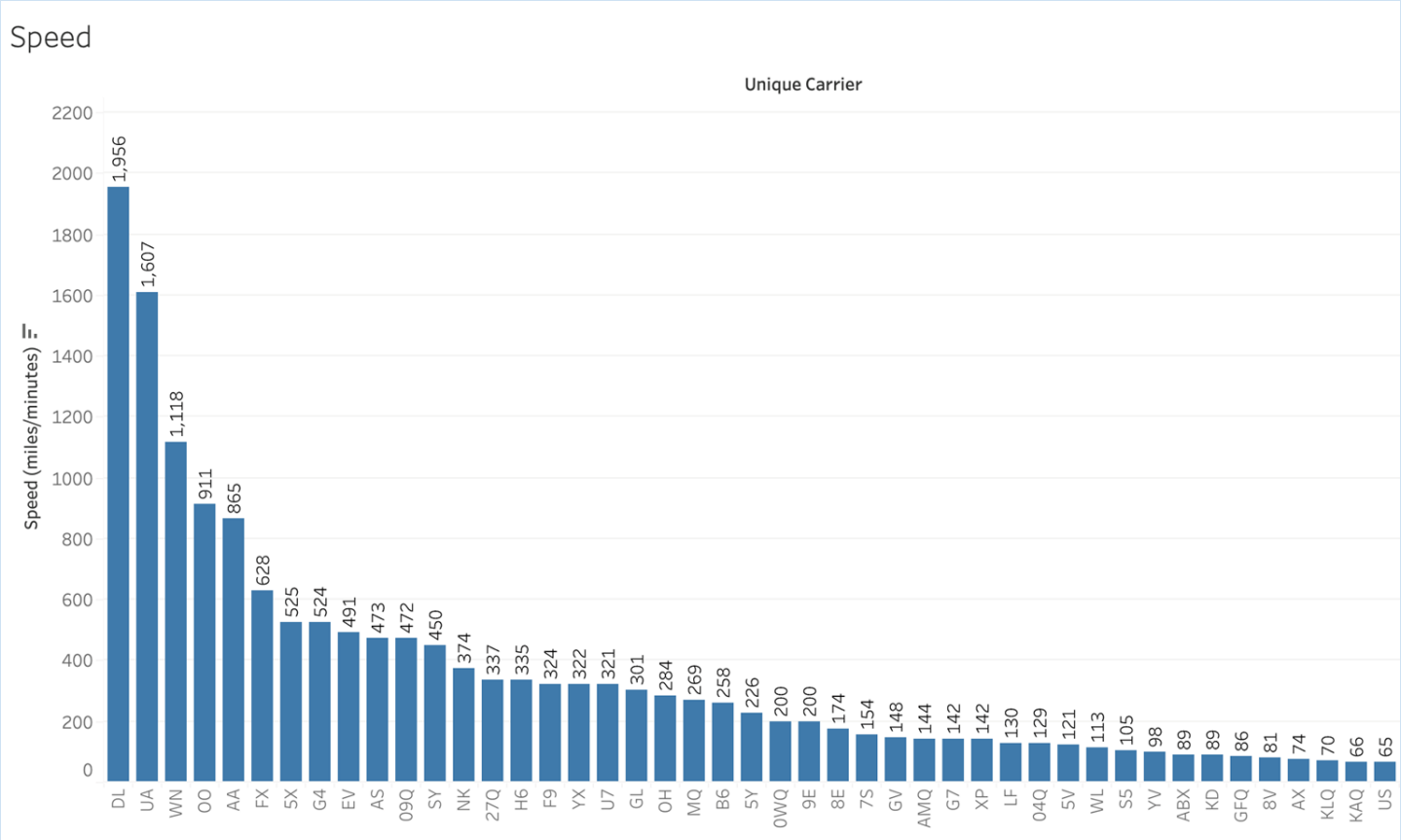
Passenger (Month)



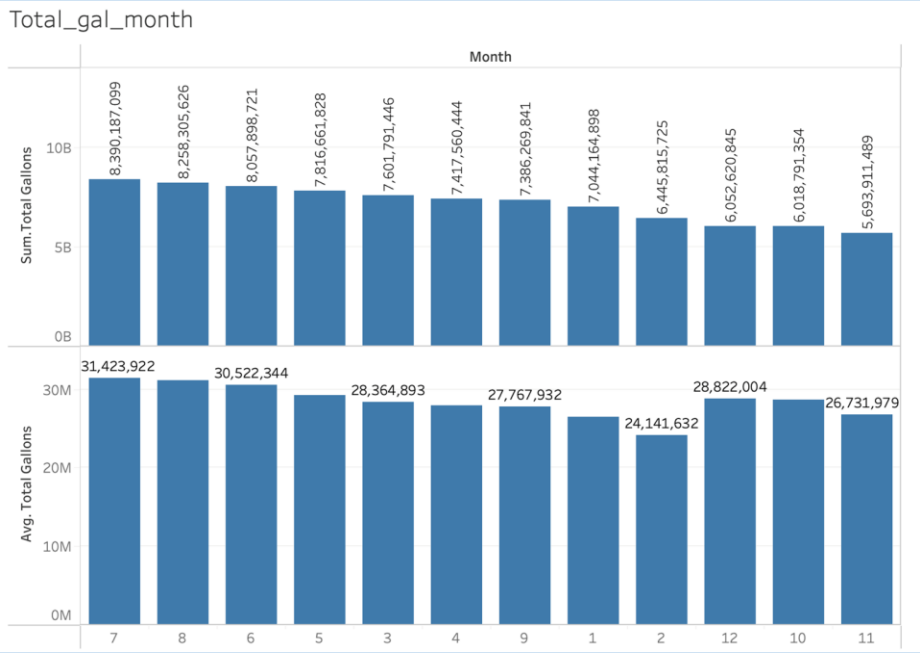
Passenger (Carrier)

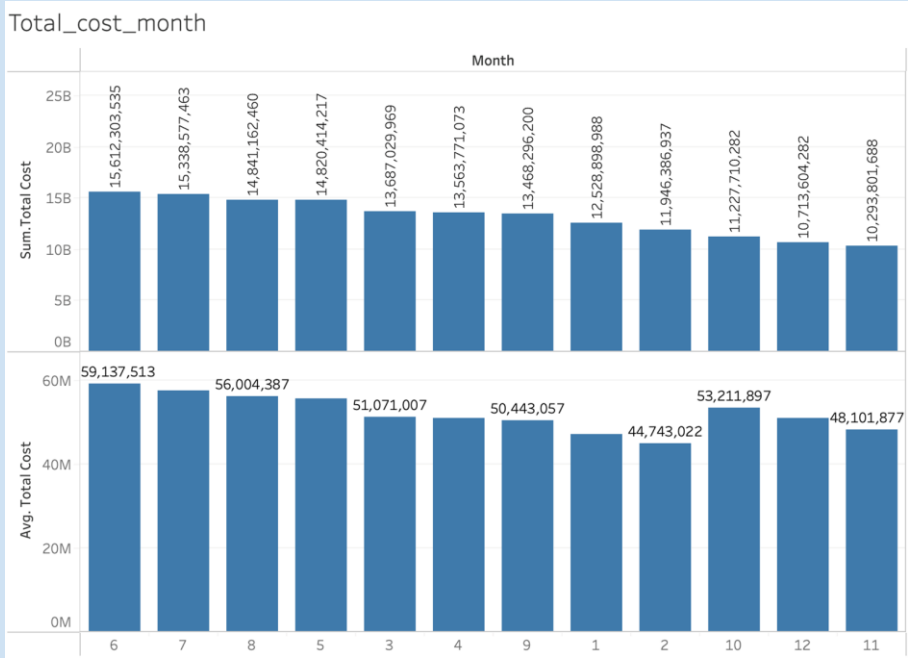


Speed (Distance/Air time)

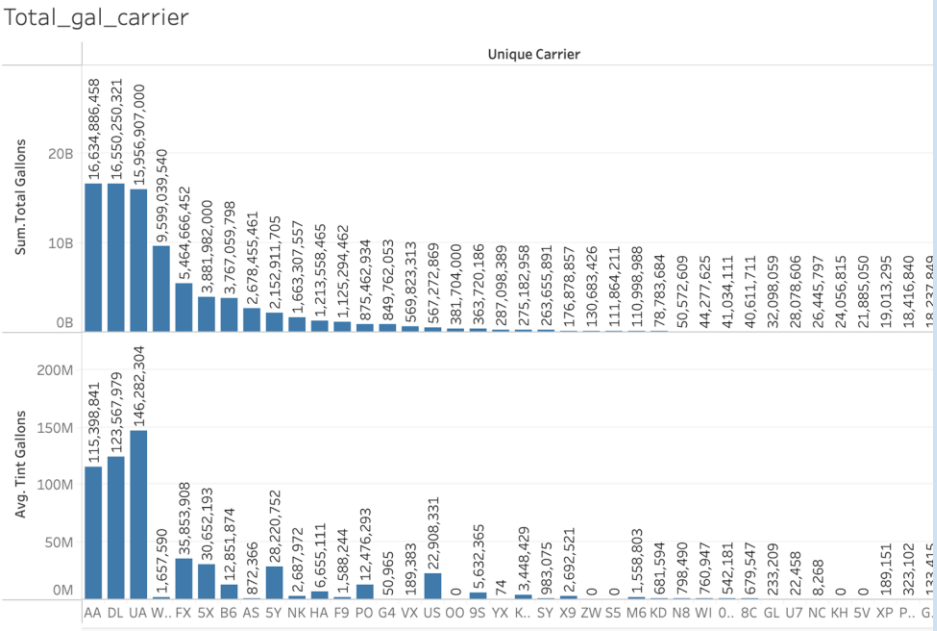


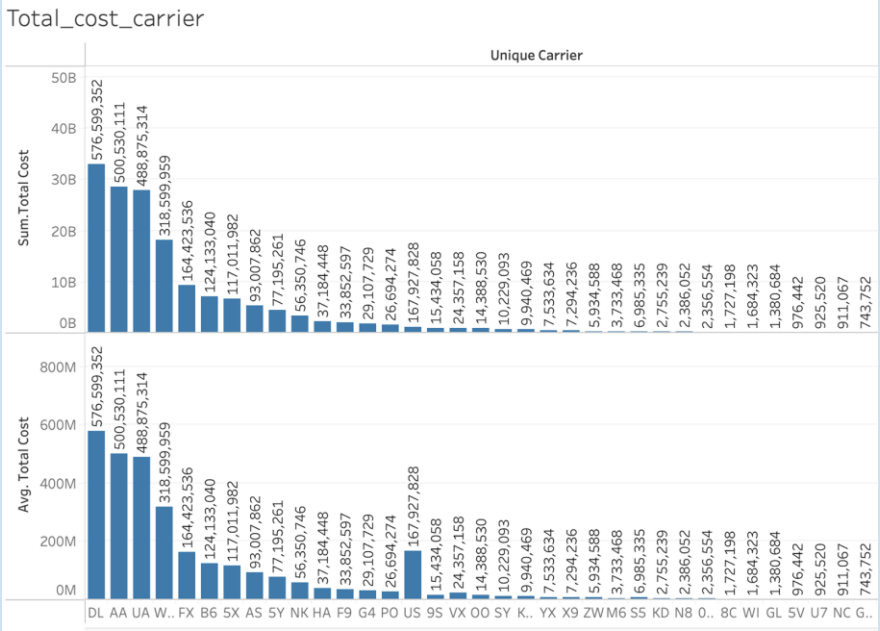
Fuel Consumption (Month)



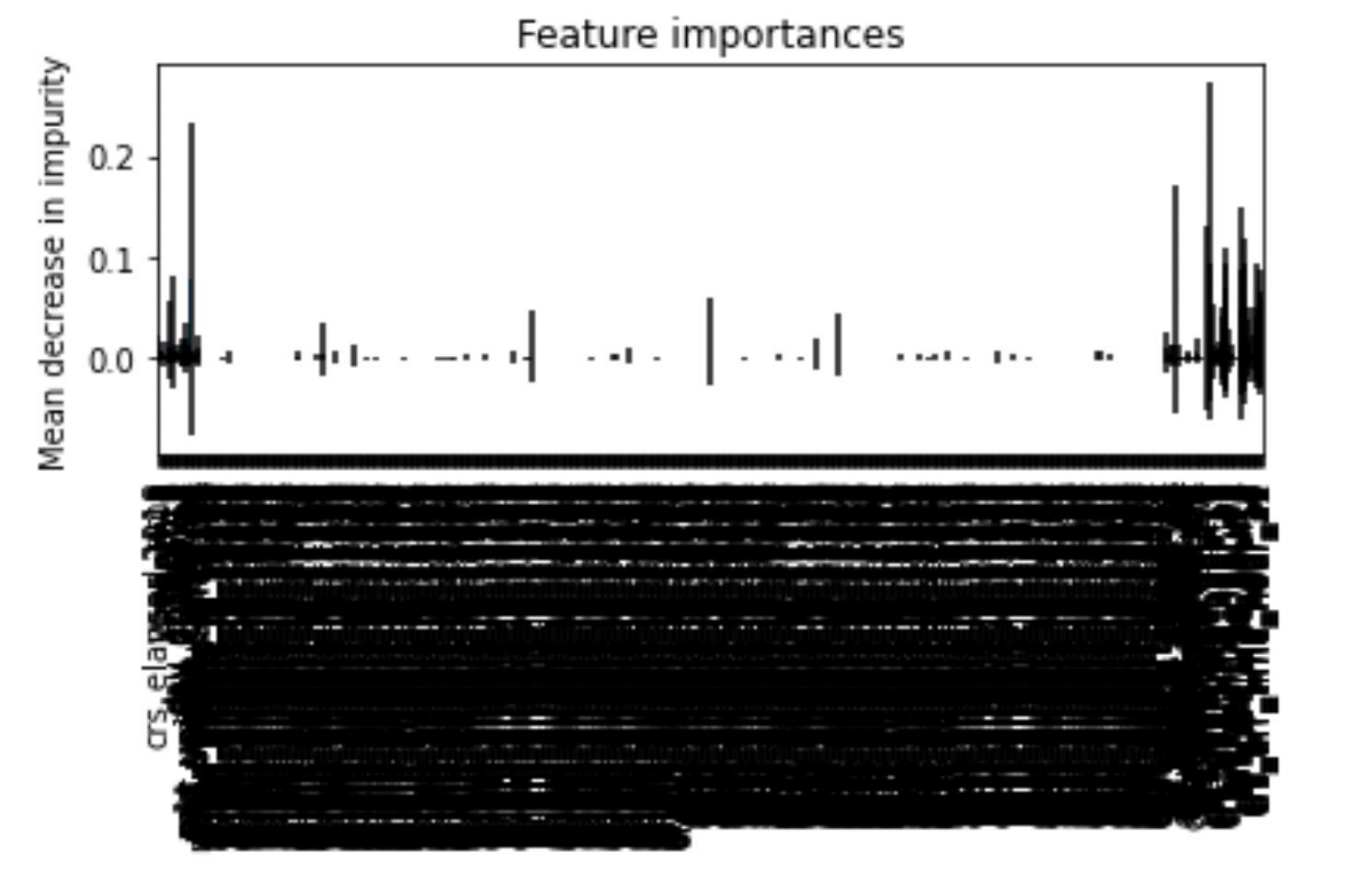


Fuel Consumption (Carrier)





Feature importance (???)



03

Data Preparation & Feature Engineering

What Impacts the Flight Delays?

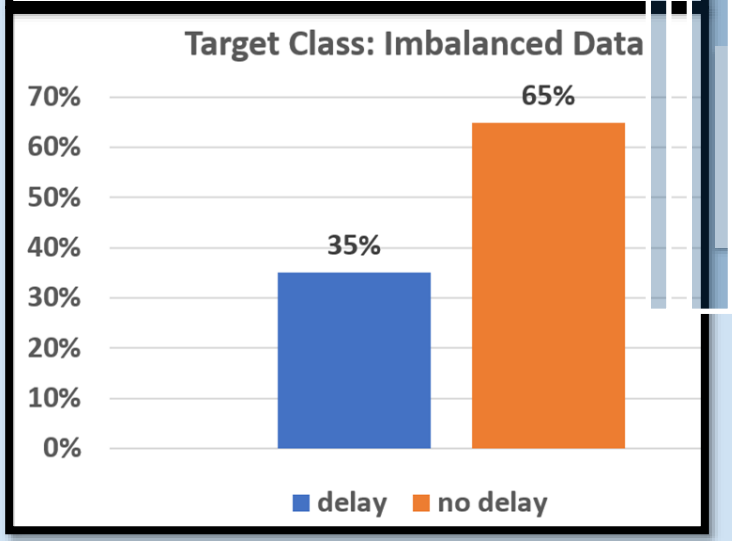
- Average measures from Origin to Destination
- Speed of the aircraft
- Traffic at the airports
- Seasonal (Time of Year / Month)
- Hours of flight operation
- Unique carriers
- Unpredictable (weather, etc)

Data Preparation: Flights

SQL Database → 1 → Pandas DataFrame

2

	Rows	Data Size
delay	5,476 K	1,114 MB
no delay	10,140 K	1,973 MB
total	**15,616 K**	**3,087 MB**



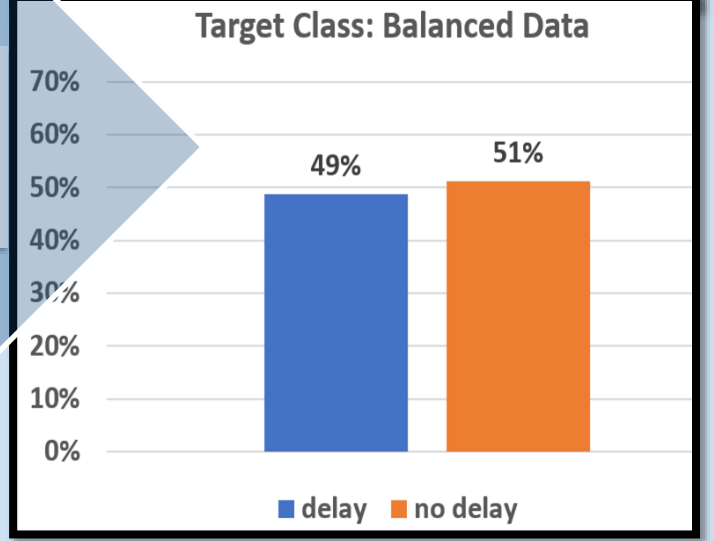
3

- 2019 delay data
- 2019 no delay data
- 2018 delay data
- 2018 no delay data

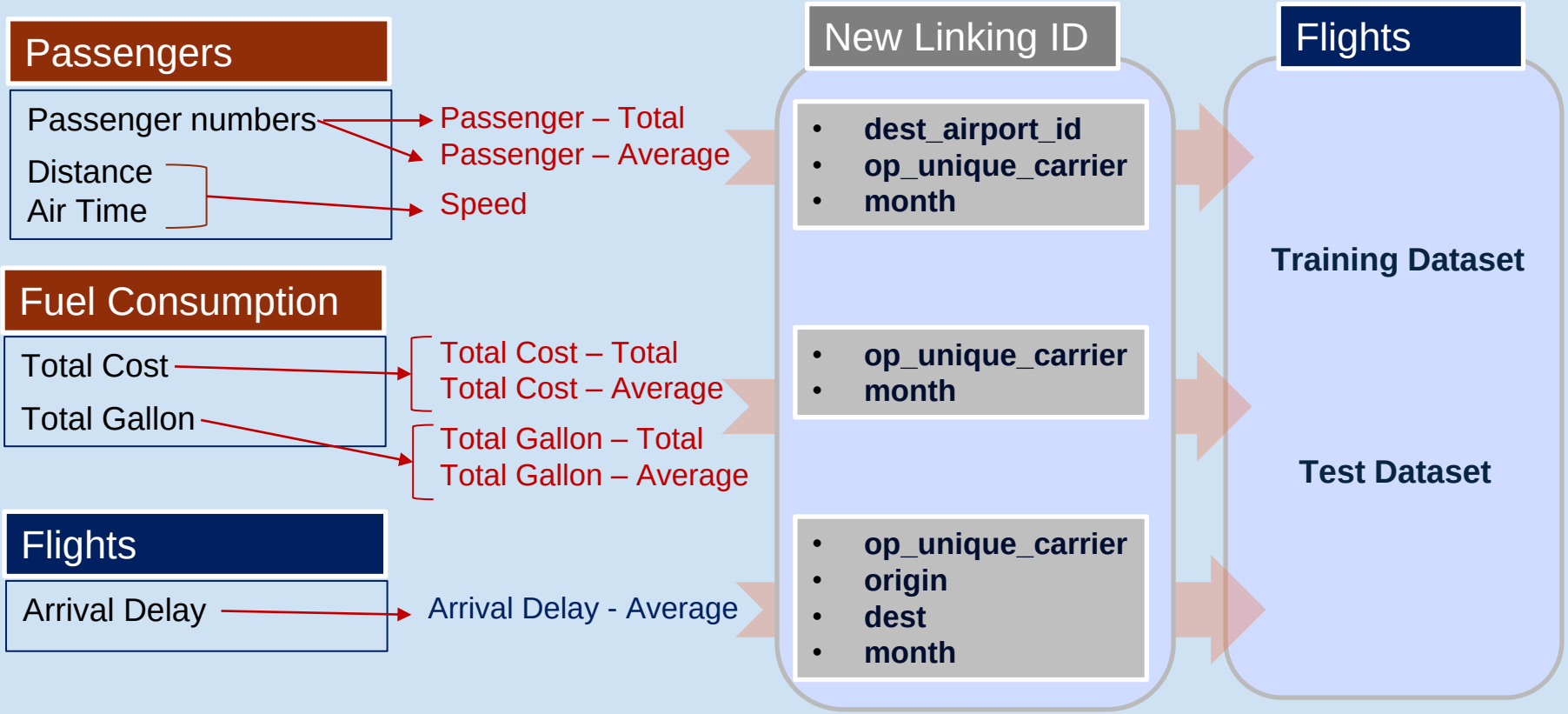
4

- December & January, each year
- Random resampling / balanced
- 5 datasets with different data size (9K, 18K, 56K, 167K & 2.6M Rows)

	Rows	Data Size
delay	27 K	4.2 MB
no delay	29 K	4.4 MB
total	**56 K**	**8.6 MB**



External Data



Feature Engineering

- CRS Elapsed Time
- Distance
- Unique Carrier Flight Number
- Total Flight Numbers (Origin / Destination)

- Speed
- Average & Total Passengers
- Average & Total Fuel / Costs
- Average Arrival Delay

Statistical Variables

- Origin Airport
- Destination Airport

- Tail Number
- Origin City Name
- Destination City Name

Ordinal Variables

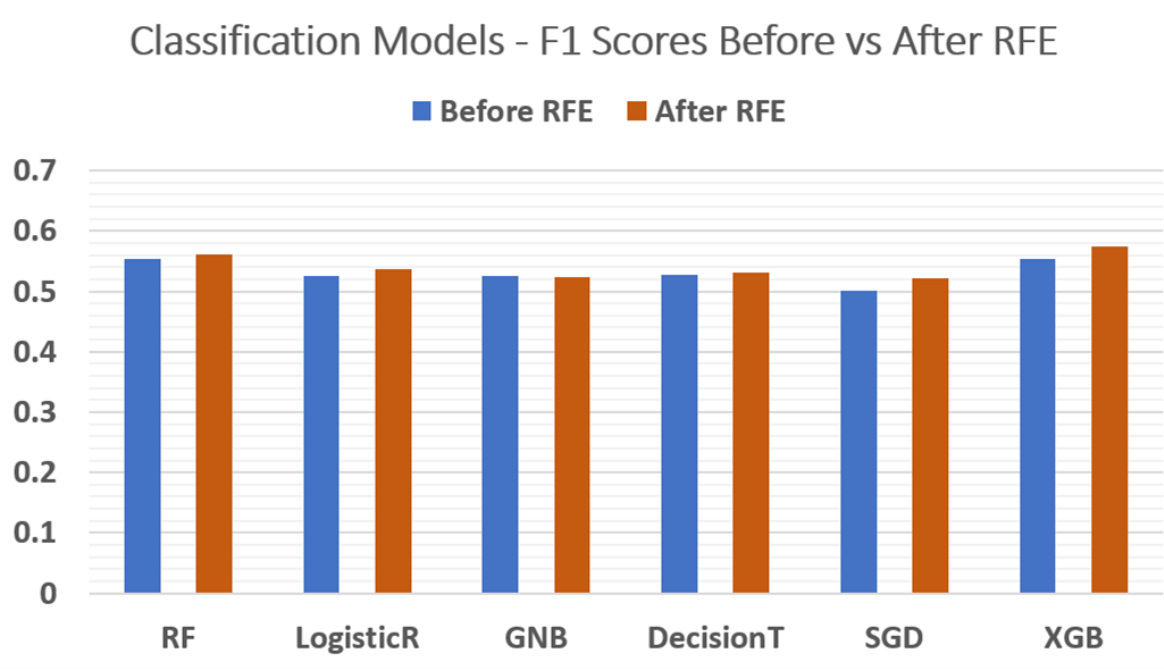
- Year
- Month (January & December)
- Day of Week
- Unique Carriers (Mkt & Op)

- Hours
- Airport Locations (Origin & Destination)

Dummy Variables

High Dimensions → *Complex Models*

Feature Engineering - RFE



RFE (estimator = DecisionTreeClassifier())

Column: 0, Selected True, Rank: 1.000
Column: 9, Selected True, Rank: 1.000
Column: 10, Selected False, Rank: 21.000
Column: 11, Selected True, Rank: 1.000
Column: 12, Selected True, Rank: 1.000
Column: 13, Selected True, Rank: 1.000
Column: 14, Selected True, Rank: 1.000
Column: 17, Selected False, Rank: 14.000
Column: 18, Selected False, Rank: 16.000
Column: 19, Selected True, Rank: 1.000
Column: 23, Selected False, Rank: 5.000
Column: 24, Selected False, Rank: 22.000
Column: 25, Selected True, Rank: 1.000

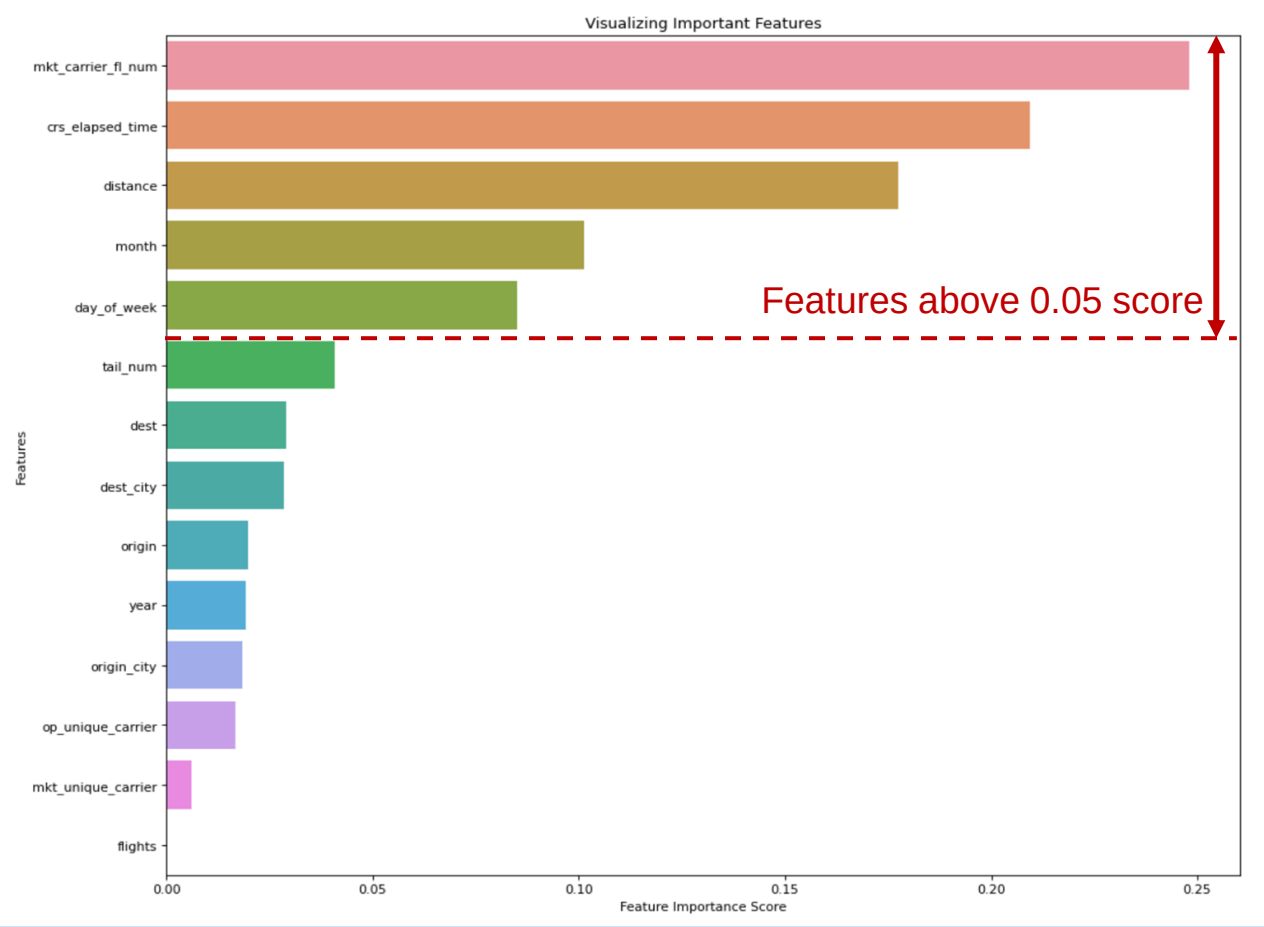
Column: 47, Selected True, Rank: 1.000
Column: 48, Selected False, Rank: 7.000
Column: 49, Selected True, Rank: 1.000
Column: 50, Selected True, Rank: 1.000

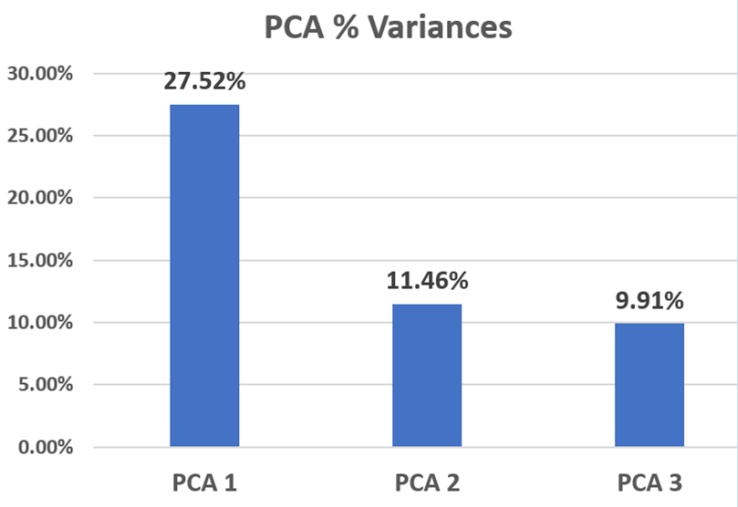
- **RFE improved the Classification Model Performance by 2% to 4%.**

RF Feature Importance

```
RandomForestClassifier(random_state=0)

mkt_carrier_fl_num     0.248034
crs_elapsed_time       0.209388
distance               0.177535
month                  0.101220
day_of_week            0.085099
tail_num               0.040926
dest                   0.029121
dest_city              0.028552
origin                 0.019868
year                   0.019130
origin_city            0.018260
op_unique_carrier      0.016707
mkt_unique_carrier     0.006160
flights                0.000000
```





Regression - PCA

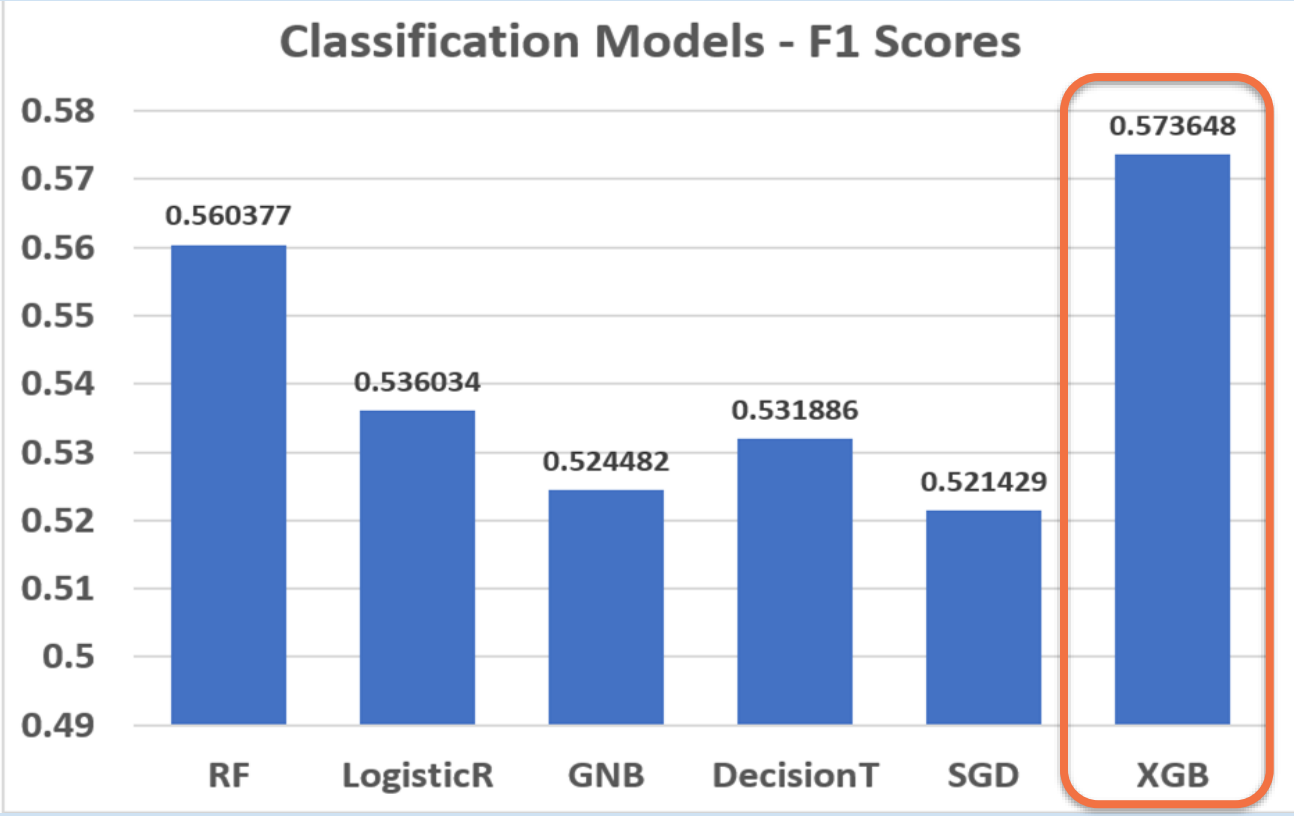
- **PCA 1:** Average Fuel Cost
- **PCA 2:** Average Passenger Numbers
- **PCA 3:** CRS Departure Time

```
Top 3 most important features in each component
===============================================
Component 0: ['total_cost:mean_fuel', 'total_gallons:mean_fuel', 'total_cost:sum_fuel']
Component 1: ['mean_passengers', 'sum_passengers', 'month_12']
Component 2: ['crs_dep_time', 'crs_arr_time', 'dest']
Component 3: ['crs_elapsed_time', 'distance', 'total_cost:sum_fuel']
Component 4: ['dest', 'origin', 'speed_passengers']
Component 5: ['speed_passengers', 'origin', 'avg_arr_delay']
Component 6: ['avg_arr_delay', 'speed_passengers', 'dest']
Component 7: ['mkt_carrier_fl_num', 'dest', 'origin']
Component 8: ['mkt_carrier_fl_num', 'dest', 'origin']
Component 9: ['year_2018', 'year_2019', 'mkt_carrier_fl_num']
Component 10: ['mkt_unique_carrier_AA', 'op_unique_carrier_AA', 'sum_passengers']
```

04

Machine Learning Model & Performance

Classification Model Performance



XGBoost was the highest performing machine learning model.

*Score ranged from 0.57-0.61

Regression – Change in R2 Scores

	linear_reg	ridge	lasso	decision_reg	random_forest	knn	xgboost
R2	0.023922	0.023923	0.000901	-0.991440	-0.013324	-0.119903	0.031551

	linear_reg	ridge	lasso	decision_reg	random_forest	knn	xgboost
R2	-2.564856e+15	0.004680	0.000914	-0.804665	-0.091168	-0.122313	-0.049703

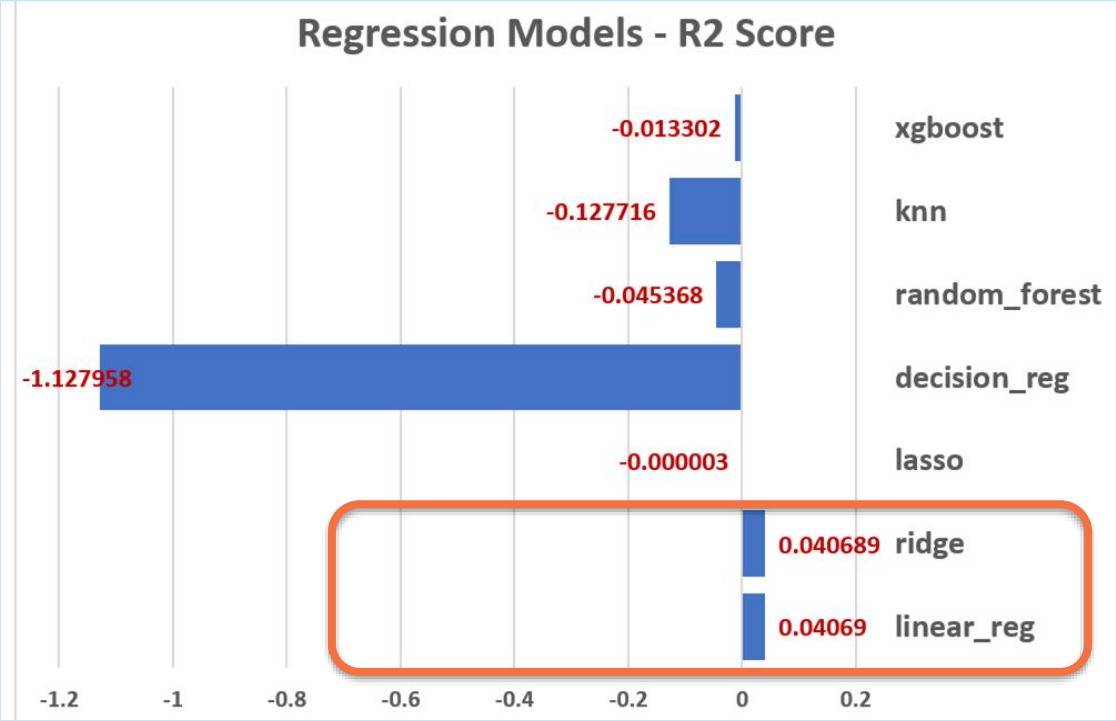
	linear_reg	ridge	lasso	decision_reg	random_forest	knn	xgboost
R2	0.040690	0.040689	-0.000003	-1.127958	-0.045368	-0.127716	-0.013302

- 117 columns : 5 Dummy & 5 Ordinal Variables
- No PCA

- 850 columns : 10 Dummy Variables
- No PCA

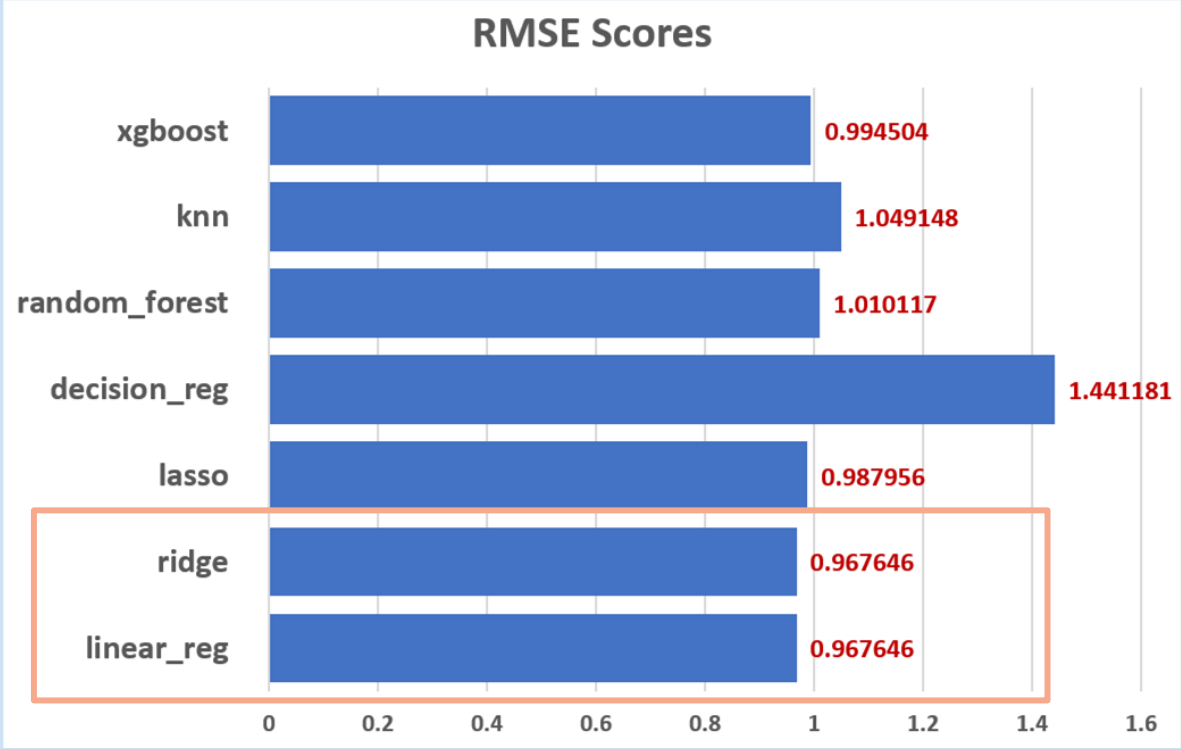
- 65 columns : 5 Dummy Variables
- PCA

Regression Performance: R2



*** Linear & Ridge Regression Models had the higher R2 scores at 4.1%, where as other models had all negative R2 scores.**

Regression Performance: RMSE



*** Linear & Ridge Regression Models had the lowest RMSE scores at 0.97.**

Hyperparameter Tuning

Building the pipeline for classification and regression models

Was not run...

GridSearchCV

XGB Classifier

- Max depth = 3 to 10
- Learning rate = 0.01 to 0.1
- Cosample bytree = 0.3 to 0.7
- n_estimators = 50, 100, 500

- **F1 Score**
- **Accuracy**
- **Confusion Matrix**

Was run!

RandomizedSearchCV

XGB Regressor

- Max depth = **3** to 20
- Learning rate = **0.01** to 0.3
- Cosample bytree = 0.4, **0.6**, 1.0
- n_estimators = 50, 100, **500**

- **R2:** **4.1%** *from -0.01%*
- **MAE:** **0.49** *from 0.50*
- **MSE:** **0.93** *from 0.99*
- **RMSE:** **0.96** *from 0.99*

Summary

Flight delay is highly impacted by

- Seasonality
- Distance
- Number of Passengers
- Fuel Usage / Costs
- Carrier Flight Number

Variability and unprecedented events also impact the flight delays, which make the prediction highly unpredictable.

Challenges

Majority of time allocated on data preprocessing and restructuring data structures, and not enough time on model implementation stage

Creating complex models by increasing dimensionality

Extremely slow for Hyperparameter Tuning and for sophisticated machine learning models

Limited computer capacity & RAM issue

THANKS

Do you have any questions?