CAPSTONE PROJECT

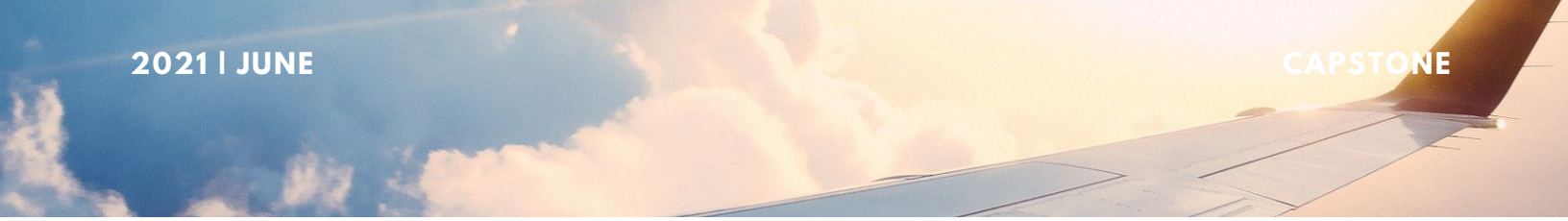
FLIGHT ESTIMATOR

A FLIGHT PRICE PREDCTION DATA SCIENCE PROJECT

Brought to you by
Peace Zigah

2021 | JUNE

CAPSTONE PROJECT



INTRODUCTION

Flights are a significant consideration in a travellers' budget, representing up to 50% of international travel costs1. COVID-19 has also sent shockwaves throughout the travel industry, making it more difficult for consumers to budget appropriately for vacations. Given these trends, this Capstone investigates if Machine Learning models and techniques could be used to predict flight prices. The MVP of this project is to train a machine learning model to predict prices; however, the end goal is to create a web-based application for travellers to look up their desired itinerary and see the estimated price of a flight (airline agnostic). Flight price aggregators like Google Flights and Hopper exist; however, they are airline-dependent and often are not ideal for long-range trip planning. Data acquisition is first required to begin tackling this business problem.

DATA SOURCE

For this project, three different data sources were used to create the data frame used for modelling.

United States Department of Transportation - Bureau of Transportation Statistics
Representing the core of my data set, Data Bank 28 Market Data is a domestic and international traffic data report which includes U.S. and foreign airline carriers. The report features U.S. direct airline flights for 2018-2019 by quarter with the following data attributes:

- YEAR: The year the data occurred
- QUARTER: The three month period of time the data occurred
- ORIGIN: The departure airport of the flight world area code
- ORGIN_STATE_NM: The departure airport state
- DEST: The destination airport for the flight world area code
- DEST_STATE_NM: The destination airport state
- PASSENGERS: The number of passengers per booking
- MARKET_FARE: The total airfare cost of the booking $ U.S.
- MARKET_DISTANCE: The distance in miles between the origin and destination airport
- DISTANCE_GROUP: Indicates which airports are within 500 miles of each other

Federal Aviation Administration
This department provided the Airport Operations and Ranking Reports which included data on the flight volume that arrived at U.S. airports by month. The full data attributes included:

- QTR_YEAR: The quarter and year of the data
- Date The month and year of the data
- Facility: The airport of the flight world area code
- State: The state of the airport
- Total Operations: The flight arrival volume for the airport

U.S. Energy Information Administration (EIA)
The EIA is responsible for collecting, analyzing, and disseminating U.S. energy information to promote the public understanding of energy and its interaction with the economy and the environment. This data source provided access to Monthly Retail Gasoline and Diesel Prices which included the following data attributes:

- QTR_YEAR: The quarter and year of the data
- DATE: The year and month of the data
- GAS_PRICES: The monthly average retail gap price in $ U.S.

Once the data sources were acquired, Exploratory Data Analysis could begin.

EDA

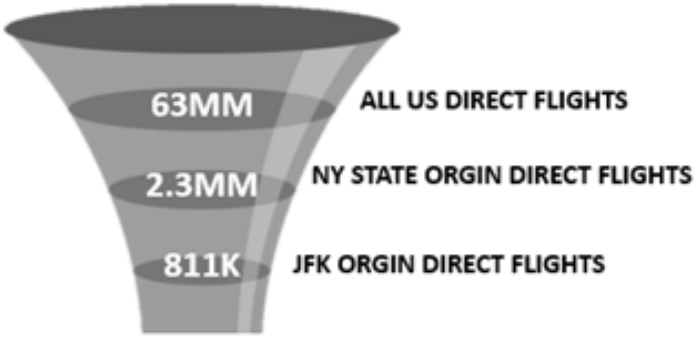
Due to a large amount of data within this dataset, Amazon Sagemaker and s3 were used. In summary, the below steps were taken in the data cleaning and preprocessing stage.

DATA CLEANING ACTION PLAN:

- For the gas data frame, the average gas price was calculated by QTR_YEAR using the groupby() and .agg(mean) to align with the core dataset.
- The column names for updated to align the naming convention for the data.
- The carr_load data frame had date gaps in the Facility column; as a result, the average flight volume for each destination was used instead. This was achieved by using groupby() on Facility and .agg(mean).
- The flights .csv's were merged into one data frame.
- PASSENGERS/ MARKET_DISTANCE converted into integers to save memory space.
- QUARTER / YEAR was converted into a string to prepare the column to have a character added.
- A new column was made called QTR_YEAR and QUARTER and YEAR was dropped as they are no longer required.
- Another column was created that calculates the flight price per passenger. It is called AVG_FLIGHT_PRICE as MARKET_FARE is the total cost of the booking, which may include several passengers.
- All of the data frames were merged using concat() to simplify the data cleaning process.
- DISTANCE_GROUP, Unnamed: 10, Unnamed: 9 and DEST_STATE_ABR were dropped from the data frame as they are not required for the analysis
- AVG_FLIGHT_PRICE and QTR_YEAR were moved to the front of the data frame to support modelling and visualizations.
- A Null_checker function created to check for nulls within the data. Reviewing the nulls, 11.67% of the DEST_STATE_NM are null values. To address this issue, the mode of the column was inputted.
- Categorical data fields were converted into numerical values using get_dummies and were appended to the data frame to support model performance.

DATA FUNNEL

There are 63MM observations within this dataset, making it challenging to analyze the data. As a result, the project's scope was narrowed to focus on flights originating from John F. Kennedy (JFK) airport in New York City. JFK is one of the busiest airports in the United States and reduced the record count to a more manageable 810,601. The data is now in a position to be visualized.

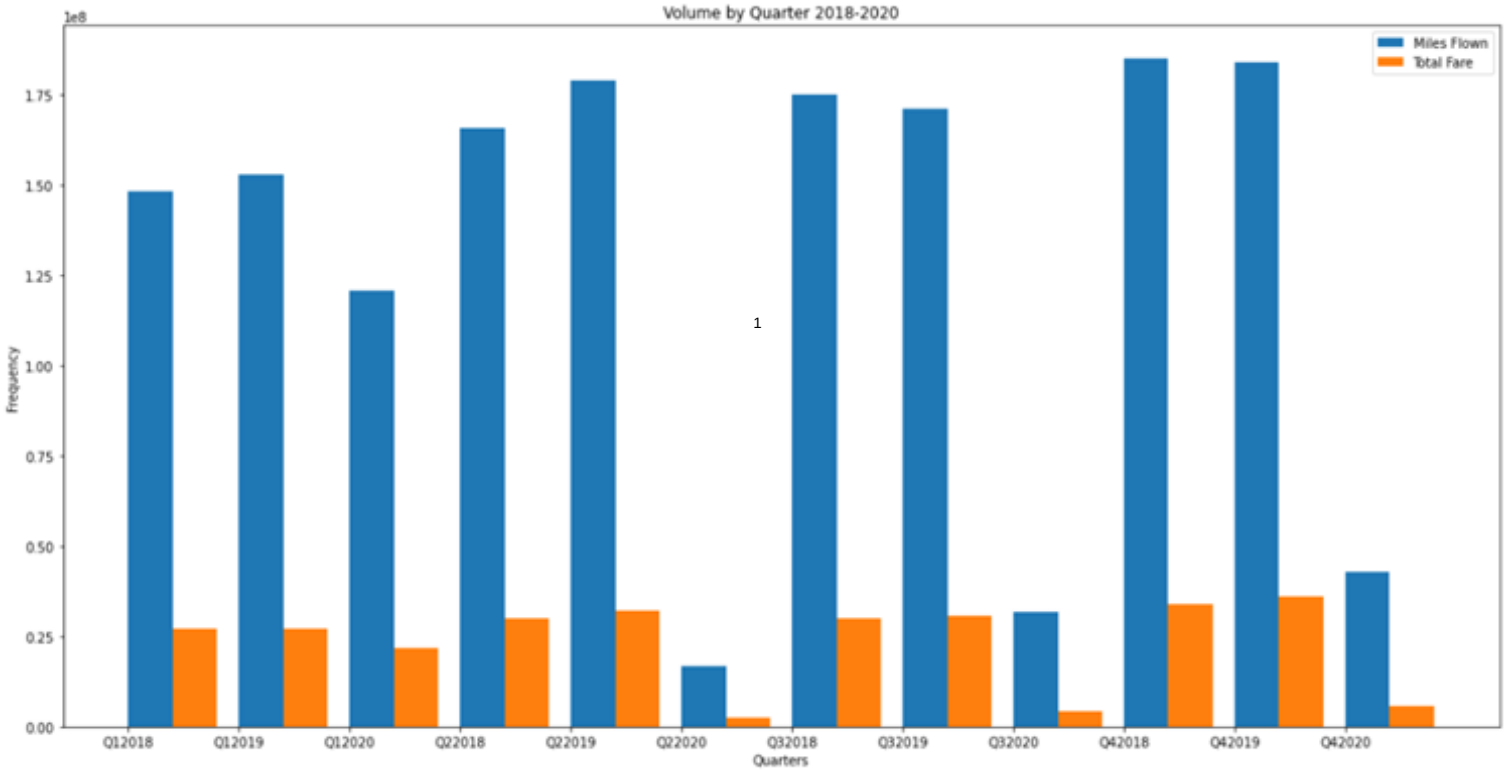


DATA VISUALIZATIONS

Within the Notebook, there were several visualizations made to learn about the data and its behaviour. For this report, visualizations that lead to data parameter changes are highlighted.

REMOVAL OF 2020

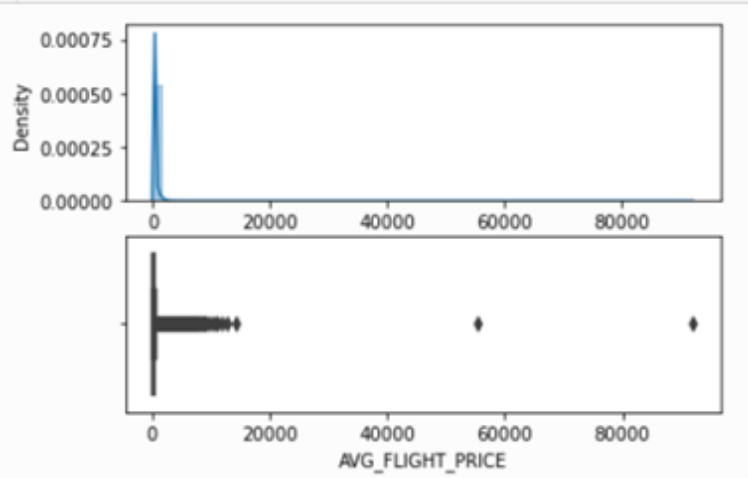
Originally in EDA, the scope of this project included data from 2018 – 2020. As modelling began, the results were poor, with R2 on average 21% lower than the final model result. The graph below illustrates why removing 2020 had such a significant impact. This visualization shows the total volume of miles flown and money spent by quarter for this data set. Naturally, 2018 and 2019 had similar activity. However, 2020 is very different. For example, Q2 Year-over-Year (YoY) had a -88% decline in miles flown. This type of drop impacted the flight prices dramatically in 2020, making it more difficult for the model to make accurate predictions, which is why 2020 was removed from the analysis.



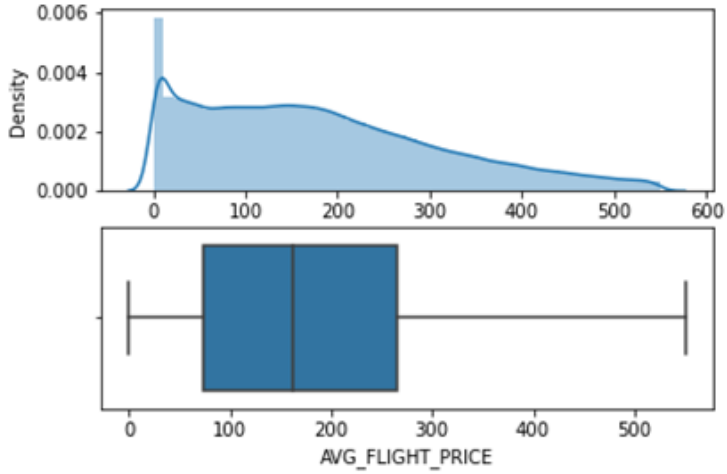
OUTLIERS

There was an extensive range in flight prices within the data set ranging from 0 - $100K per person. However, most of the price points were $0 - $550, as showcased by the Distribution and Box Plot graphs below. As a result, 65,163 outliers are within the data (price points above $550), representing 8% of the data. The model was re-run removing these outliers, and the model performance on average increased on average 14%. As a result, the data parameters were adjusted to remove outliers as displayed by the Outliers Removed graphs.

WITH OUTLIERS:



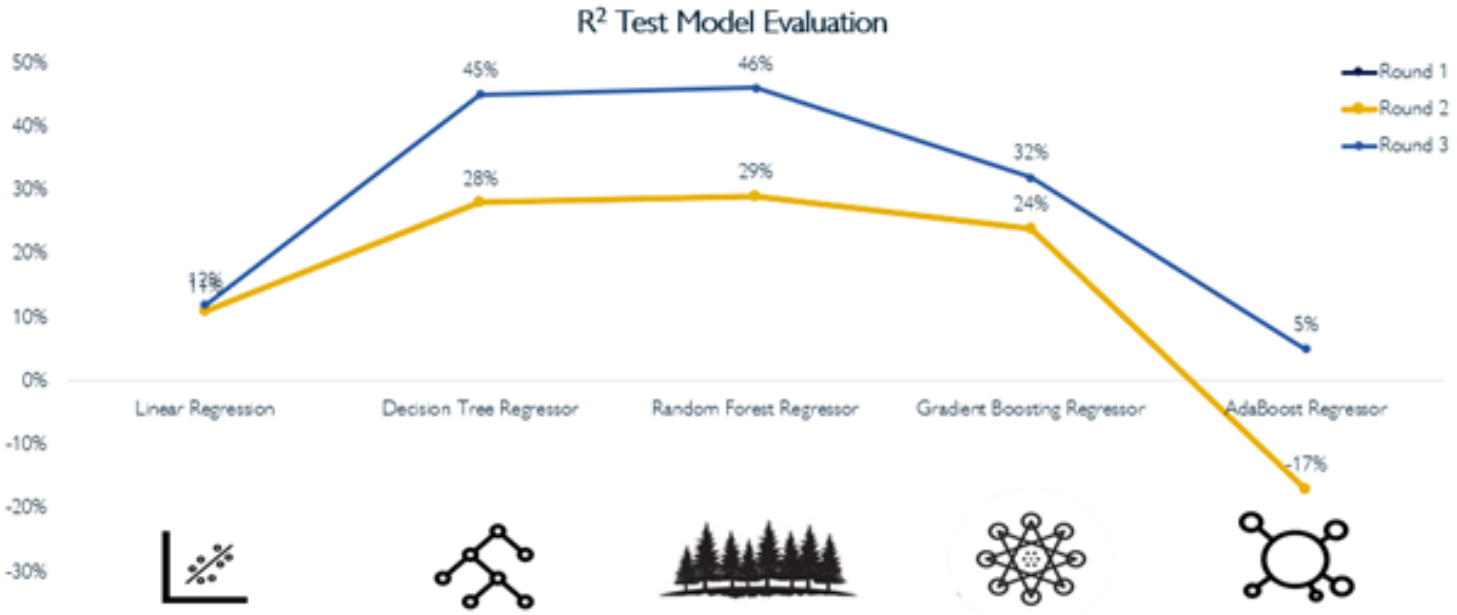
OUTLIERS REMOVED:



MODELLING

As eluded to above, there were several rounds of modelling. Twenty percent of the data was test data, and a random_state leveraged to maintain similar results with each iteration of the model runs. A function was created to support running models. This function fit the data and calculated model performance using R-squared (R2), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). These metrics are all statistical measurements appropriate for using regression models to measure the difference between the test and predicted values. R-squared was referred to most when analyzing performance. The function also included a distribution plot to visualize the delta between the predicted values and the test data (a.k.a - error). Since the defined dependant variable is continuous, supervised regression machine learning models were used.

Due to the data size, Round 1 of modelling was first run using the default hyperparameters. The models took several hours to run, so PCA was performed, which decreased the data frame's dimensionality by 23% and the models re-run with nearly identical performance (Round 2). Round 3 of modelling removed outliers and was by far the best performing model. The below models were attempted. Randomized Grid Search was also attempted on Round 1, and 3 did not positively impact model performance, so the default model hyperparameters were maintained.



CONCLUSIONS

Can Machine Learning models and techniques be used to predict flight prices? In this assignment, it was possible to predict 46% of the test data using the Random Forest Machine Learning model. When there is an error in the prediction, on average, it is off by $69. Various models were attempted, such as Linear Regression, Decision Tree, Random Forest, AdaBoost and Gradient Boosting with varying results. However, Random Forest consistently outperformed the other Machine Learning Models. The Scientific Process is iterative, and the data parameters needed to be re-evaluated and dimensionality reduced to save computation time and improve model performance. With the removal of 2020 and outliers, the model saw a respectable +21% and +14% increase in performance from the first model run (not shown, included 2020 data). Unfortunately, Randomized Grid Search was used for hyperparameter optimization due to the data size and did not improve performance. In the end, the best parameters were the default model parameters.

	Linear Regression	Decision Tree Regressor	Random Forest Regressor	Gradient Boosting Regressor	AdaBoost Regressor
R^2:	Train: 13% Test: 12%	Train: 51% Test: 45%	Train: 50% Test: 46%	Train: 32% Test: 32%	Train: 5% Test: 5%

The work does not end on this project as a next step; Bokeh was used to visualize that data and make it easy for the user to interact with the data and determine their flight costs. Additional data will also be added to improve the results of the model.

SOURCE:

1. https://www.valuepenguin.com/average-cost-vacation