CAPSTONE PROJECT: FINAL REPORT

Flight Plans & CO2 Emissions

By: Xenel Nazar |E: xenel.nazar@gmail.com | Date: Sept 26, 2022

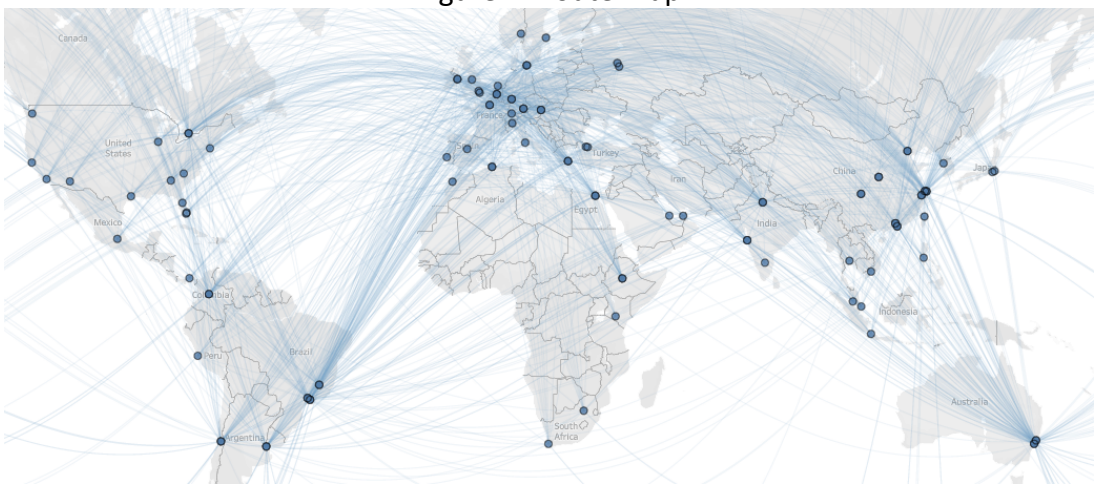
PROBLEM STATEMENT

Certain modes of transportation, such as air travel, play a vital part in the global economy in connecting people across the world, either for leisure or for business. However, the issue of climate change has impacted how many of us go about our daily lives. There have been campaigns to spread awareness, and have each person strive to limit their impact on the environment as much as they can. This capstone project looks to how everyday people can make better decisions in their trip plans, to limit their environmental impact. In addition, stakeholders for the trip plans, such as airlines, can look at the insights to see how they can improve efficiencies in their operations, that can also help with the bottom line.

DATA OVERVIEW

The data used for this capstone project was scraped by Barking Data a web mining service firm, from Google Flights, an online flight booking service run by Google/Alphabet. The data scraped by Barking Data was hosted on Kaggle. The data of almost a million rows detailed trip information, on a row-by-row basis, from destination and origin airports across the world, as detailed in the route map detailed in Figure 1. The data included fare information, carbon dioxide emissions information, as well as trip information: Destination/Arrival Times, Aircraft used, Airline Operators, Trip Duration in minutes, and Number of Stops/Legs.

Figure 1: Route Map



DATA CLEANING & EXPLORATORY DATA ANALYSIS

To make the data usable for us prior to modeling, we needed to conduct various data cleaning, feature engineering, and pre-processing.

Some of the various data cleaning steps taken include: adding details and coordinates for all airports listed, removing duplicate information, imputing null values, when possible, CO2 Emissions verification, splitting aircraft type used and airline operators by # of stops listed, standardizing times to the UTC/GMT timezone, as well as calculating the distance between two airports.

In the process of data cleaning, a new feature was created, the *km per lb of CO2 emissions generated.* This helps detail how efficient the trip is, in regards to the kilometers covered per pound of CO2 emissions generated. This was then simplified to following:

- "High Efficiency/Utilization" – mapped as "1" to detail trips generating >4km per lb
- "Low Efficiency/Utilization" – mapped as "0" to detail trips generating 0-4km per lb

Some of the insights generated include: finding the positive correlation between fare prices and CO2 emissions, as well the correlation between distance covered and CO2 emissions. Cross-country trips generate the most emissions, but on average, they are more efficient than intra-country trips. Modern aircraft, like the Boeing 787, tend to be more efficient than older turbo-prop aircraft currently in service. Certain carriers, like Low-Cost-Carriers, are not only known for their low fares, but their strive to be more efficient, and the numbers do say that the industry can learn from how they operate. In addition, the most efficient number of stops one should take would be non-stop flights, followed by 2-stop trips. Lastly departure times, such as 6AM UTC, and arrival times of 3AM UTC, would lead likely lead to more efficient flights.

Pre-processing of the numerical and categorical columns was taken to transform and prepare the data for modeling.

MODELING

Prior to modeling, additional steps needed to be taken, including conducting an X & Y split to identify our target variable, which in this case is the new feature km/lb classification for high and low efficiency trips.

The data was also exposed to a train-validation-test split, which resulted in a roughly 40-30-30 split in the data.

For this capstone project the Logistic, SVM, Decision Tree, Random Forest, and XGBoost classifiers were used in the modeling of the processed data. As listed in the table in Figure 2, baseline results were generated from the models, as listed and subsequently various

parameters for the tree models (Decision Tree, Random Forest, and XGBoost) were optimized by a ML pipeline to help in improving the overall results of the models.

Figure 2: Modeling Overview

MODEL	RESULT
LOGISTIC MODEL	0.88
LOGISTIC-PCA MODEL	0.88
SVM MODEL	0.87
DECISION TREE (MAX DEPTH = 1)	0.86
RANDOM FOREST (N_ESTIMATORS = 50)	0.97
XGBOOST	0.96
DECISION TREE (MAX_DEPTH = 10, MIN_SAMPLES_LEAF = 1)	0.93
RANDOM FOREST (N_ESTIMATORS = 50, MAX_DEPTH = 10)	0.89
XGBOOST (N_ESTIMATORS = 40, MAX_DEPTH = 10)	0.97

In the end the optimized XGBoost model was selected, based on the highest score. The model helped generate insights, including understanding the most important features impacting the score, such as price, duration, destination/origin coordinates, and much more.

NEXT STEPS

Further refinements to the data and the model can be taken to help improve the prediction of the model, including adding/removing features to the final dataset used for modeling during pre-processing. Potential over/under sampling of the data, as there is classification imbalance in the target variable. Altering the target variable is also other option that can be done to help generate better results. Lastly, adjusting more hyperparameters of each model through the ML pipeline can also assist in improving our models even further.