DATA1030 Final Project Report: Forecasting Emissions Across Global Flights

Angela Yeung

Brown University

December 2023

https://github.com/angelay1006/data1030project_new/tree/main

1 Introduction

In the face of a critical environmental challenge, this project explores carbon dioxide (CO_2) emissions across global flights. The overarching question is whether we can effectively forecast the CO_2 emissions of individual flights based on factors such as ticket price and flight duration. Aviation occupies a complex position in climate discourse: international flights are not attributed to any country's emission accounts, so there are fewer incentives for reduction. Moreover, in 2022, aviation rebounded to 80% of pre-pandemic levels. Due to these reasons, this topic is highly relevant in the post pandemic context. The dataset was obtained from Kaggle, uploaded by BarkingData, and it was collected through web scraping techniques from flight tracking and airline websites. While there is existing research related to predicting CO_2 emissions with ML methods, it is important to clarify that this project was not directly based on or influenced by any specific prior study.

2 EDA

2.1 Preliminary Information

The dataset consists of 998,866 rows and 18 columns. The features are as follows: `from_airport_code`, `from_country`, `dest_airport_code`, `dest_country`, `aircraft_type`, `airline_number`, `airline_name`, `flight_number`, `departure_time`, `arrival_time`, `duration`, `stops`, `price`, `currency`, `co2_emissions`, `avg_co2_emission_for_this_route`, `co2_percentage`, and `scan_date`.

2.2 Correlation Matrix

A correlation matrix was used to display the top three continuous features. Two features dropped from the matrix were the target variable itself and the average carbon dioxide emission per row.

Feature	Value
price	0.724647
duration	0.492424
stops	0.377681

Table 1: Correlation Matrix of Overall Dataset

2.3 Information on Target Variable

The distribution of emissions was plotted using a bin count determined by the square root of the total data points. It suggests that there are some flights with extremely high emissions that are pulling the mean upwards, as well as significant variability among different flights.

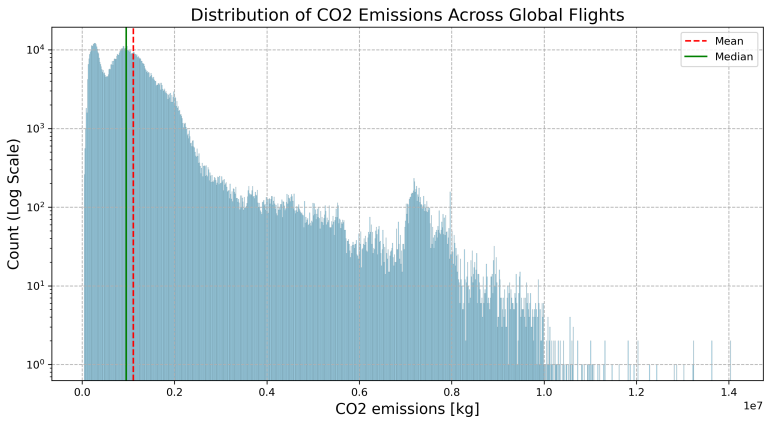

Figure 1: Distribution of emissions across flights. The mean (1111010.42) being higher than the median (956000.00) confirms that the data is positively skewed.

2.4 Continuous Variables

The `price` feature had the highest correlation with the target variable. The data points are dispersed across a range of prices and emission levels, indicating

variability. Certain areas of the graph are more densely populated, a hint at common price and emission combinations.

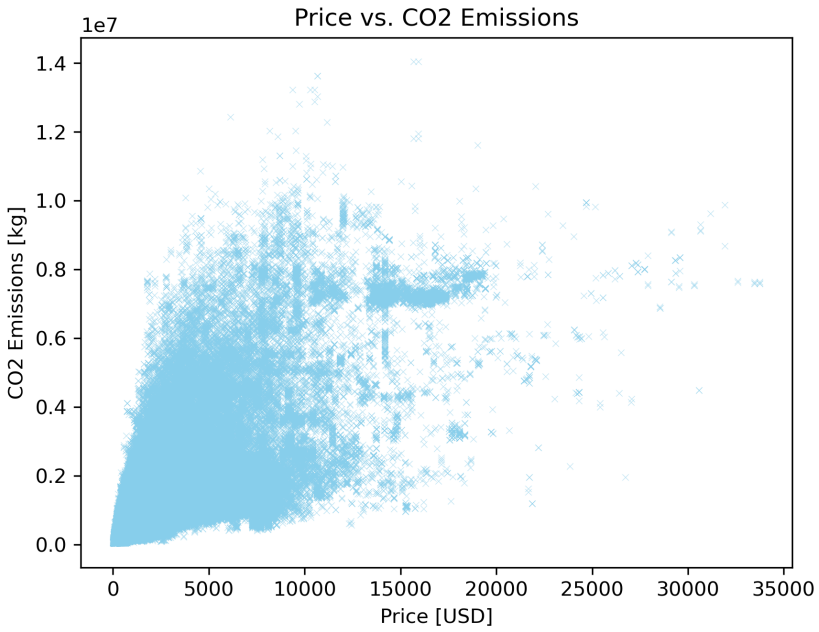

Figure 2: Scatter plot showing relationship between price and emissions.

2.5 Ordinal Variables

For the project's purposes, we can consider the `stops` feature as ordinal (see Figure 3). The visualization suggests that flights with more stops will tend to have higher emissions. The heights of the boxes show that flights with 3+ stops have a wider range of emissions - this could be attributed to distance, passenger load, or type of aircraft.

2.6 Categorical Variables

The feature `from_country` was arbitrarily chosen for visualization (see Figure 4). Since the width of the violin indicates the density of data points at that level, Australia has the highest emissions among all the countries that appear the most frequently in the dataset.

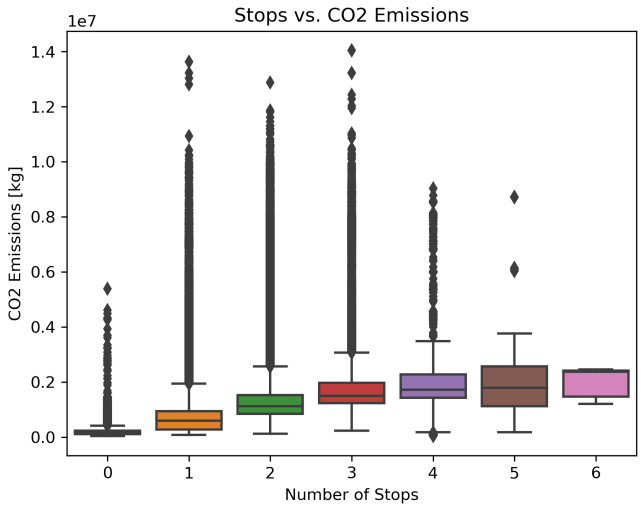

Figure 3: Box plot showing relationship between stops and emissions.

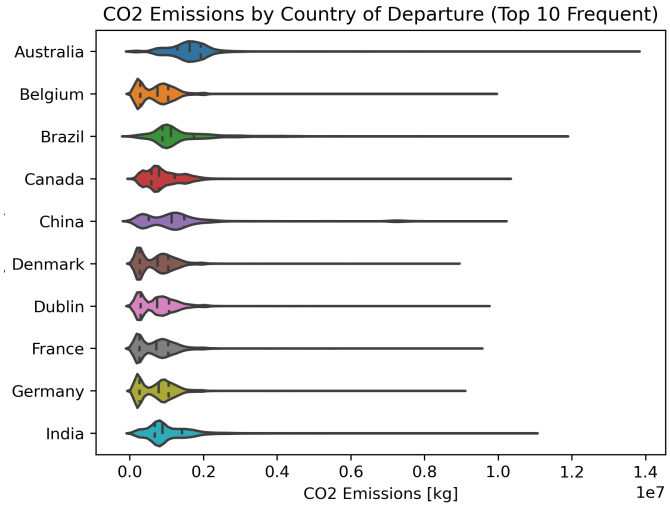

Figure 4: Box plot showing relationship between country of departure and emissions. The long "tail" for the violins observed could be due to multiple-leg flights, longer flights, or usage of aircraft with higher emissions.

2.7 Optimizing the Dataset

A key part of EDA was the decision to focus on single-leg flights; this stems from the difficulty in accurately calculating emissions for multi-leg flights. These involve various aircraft types across different legs, which complicates the attribution of emissions to specific segments. By excluding multi-leg flights, the aim is to maximize the analysis's accuracy. Since single-leg flights are more direct with consistent conditions, this allows for straightforward estimation of emissions. Rows where emission values were missing were removed, ensuring that models are only trained on complete data. Columns that were either too closely linked to the target variable or irrelevant to the analysis were eliminated, such as `co2_percentage`, `avg_co2_emission_for_this_route`, and various flight/airline identifiers. After optimization, the shape of the dataset was `(16819, 10)`. Furthermore, missing values were identified. First standardized through placeholders, followed by calculation of the total number of the missing values across the dataset, the 847 missing values in the price column accounted for 5.04% of the rows in that feature.

3 Methods

3.1 Consideration of iid/non-iid dataset

It is evident that the dataset is non-iid (not independent and identically distributed), due to time-based attributes like `departure_time`. Flights closer in time often exhibit more similarities, and there is also a geographical aspect, like flights from the same airports having common characteristics. Additionally, variables such as 'price' and 'duration' are likely correlated. My strategy for splitting the dataset revolves around the 'price' attribute for several reasons:

1. The correlation matrix suggests a link between flight prices and emissions, likely because longer, more expensive flights emit more CO2.
2. Stratifying by price ensures that each subset of data is representative.
3. This approach helps avoid bias in data splits to prevent over-representation of certain price ranges in any subset.

3.2 Splitting Strategy

Initially, time-based features were converted to numerical formats through a manual function. Recognizing the significance of the `price` variable, especially given the missing values, temporary imputation was performed using 0 to facilitate the creation of price bins, as this number does not naturally occur under this feature and makes it easy to identify for later iterative imputation. The necessity for this is due to avoiding `TypeError` when defining price bins for stratification. More specifically, this error was encountered when attempting to stratify the data based on `price`, so the missing values were temporarily imputed to ensure

that all the entries were numeric and compatible. The `train_test_split` was used twice to divide the dataset into training (60%), cross validation (20%) and test (20%) sets. After splitting, the size of each dataset was 13455, 1682, 1682, respectively.

3.3 Preprocessing and Pipeline

The next step involved applying the `IterativeImputer` distinctly to each data subset based on whether the row has 0 as a value for `price`. To address categorical features, a pipeline was established, which combined `SimpleImputer` with `OneHotEncoder`. For numerical features, a placeholder `IterativeImputer` was used, along with `StandardScaler` for normalization. Using a `ColumnTransformer`, it was ensured that each feature type went through the appropriate preprocessing steps. The expanded shape of the preprocessed data was now `(16819, 213)`.

3.4 Evaluation Metrics

The choice of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) as evaluation metrics for the baseline and actual models is justified by their ability to assess the models' performance in a regression problem. Mostly, the analysis will be focused on RMSE and R^2, as RMSE is more interpretable (same units as the target variable) and measures the average magnitude of the errors between the predicted and true values. In the context of this project, it will provide a clear metric of how much the model's predictions deviate from the observed emissions data, which is pertinent to assessing practical utility. Conversely, R^2 will show how well the model fits the data, also translating to the value of the model when it comes to stakeholders looking to estimate emissions from flight data.

3.5 ML Algorithms

Linear regression was implemented with L1 and L2 regularization for linear models; for nonlinear models, Random Forest, XGBoost, and SVM were implemented.

3.5.1 Linear Algorithms

Here, the focus was on optimizing the regularization strength of linear models (Lasso and Ridge), while also including a standard Linear Regression model. For L1 and L2, various alpha values `(10, 15, 20, 30, 50, 100)` were experimented with. Such values were chosen to cover a broad range to help in assessing impact on performance. Cross-validation was then used to calculate mean and standard deviation for MSE, RMSE, and R^2 score. The standard deviations for each metric were also calculated here. Using cross-validation was essential in mitigating the variability in performance. The usage of k-fold cross-validation

allows the models to be trained and validated on multiple data subsets, which provides a more powerful assessment. It was observed that MSE, RMSE, and R^2 fluctuated depending on the data subset used.

3.5.2 Nonlinear Algorithms

For Random Forest, a range of parameters (specifically n_estimators, max_depth) were tuned in order to identify an optimal configuration. The values were determined to be 60 estimators with a max depth of 20; these were selected based on performance in terms of the R^2 score. This fine-tuned model was then incorporated into the pipeline.

XGBoost Regression was the second non-linear approach. Similarly, the focus was on tuning the hyperparameters n_estimators, max_depth, and learning_rate. The chosen values were 225 estimators, a max depth of 5, and a learning rate of 0.2, similarly optimized in the context of the R^2 score.

The final nonlinear model explored was the Support Vector Machine (SVM) regressor. The hyperparameters tuned were `C`, `gamma`, and the `kernel` type. The optimal configuration here was C with a value of `190`, gamma set to `scale`, and the kernel type set to `linear`. As before, these were chosen based on the model's R^2 score.

With Random Forest, the variability in model performance was more obvious due to inherent randomness stemming from the feature selection design for building individual trees. It is important to clarify that all models here were evaluated under a variety of random states — this provided more information regarding the stability of the model under different conditions. The standard deviations of performance metrics also served as an indicator of how consistent these models were.

4 Results

4.1 Baseline Scoring

To establish regression baselines, mean values of `y_train` were used as basic predictions for each instance in `y_test`, which allowed calculations of MSE, RMSE, and R^2.

Baseline Metric	Value
MSE	109342193118.233
RMSE	330669.311
R^2	-0.000

Table 2: Table describing baseline model performance

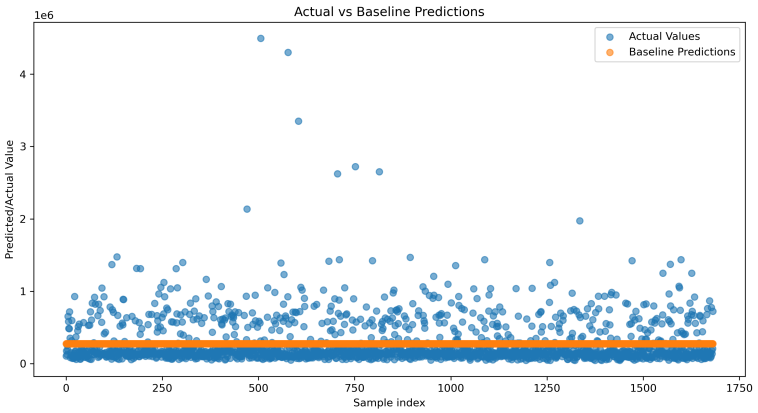

Figure 5: Scatter plot showing true values vs. baseline predictions

4.2 Best Performing Model

XGB resulted in being the best performing model, with a mean MSE of 6112169763.538, mean RMSE of 78180.367, and mean R^2 score of 0.944. This signifies high accuracy; the relatively low RMSE also suggests much more precise predictions with significantly smaller errors. However, it is important to note that the evaluation outputted identical MSE, RMSE, and R^2 scores across various iterations such that standard deviations for each metric resulted in 0.00. This is a unique case when compared to the outcomes of the other models, suggesting that this phenomenon is model-specific rather than a fundamental issue with the data setup. It points towards the consistency and stability of the XGB algorithm, but also raises concerns specifically with the model's behavior, such as its sensitivity to variations in random state.

4.3 Summary

Below are figures and results for each ML model described in the previous section. The below table shows a comprehensive overview of the four main models.

	LinearRegression	Random Forest	XGB	SVM
MSE	16093517414.237	8295257524.084	6112169763	23458334613.151
RMSE	125249.787	90546.965	78180.367	152594.891
R^2	0.812	0.906	0.944	0.735
MSE STDDEV	5270433010.187	1918814978.941	0.00	3961720747.514
RMSE STDDEV	20149.649	9823.678	0.00	13158.039
R^2 STDDEV	0.037	0.019	0.00	0.031

Table 3: Summary table of all model results

4.3.1 Linear Models

The table below provides a more detailed look into the process of generating a linear ML model.

	Linear Regression	Lasso (L1)	Ridge (L2)
MSE	16093517414.237	16405479353.001	18987752759.289
RMSE	125249.787	126585.675	136577.319
R^2	0.812	0.808	0.777
MSE STDDEV	5270433010.187	5147865657.522	5153024597.014
RMSE STDDEV	20149.649	19533.210	18286.299
R^2 STDDEV	0.037	0.035	0.031

Table 4: Summary table of linear regression model results

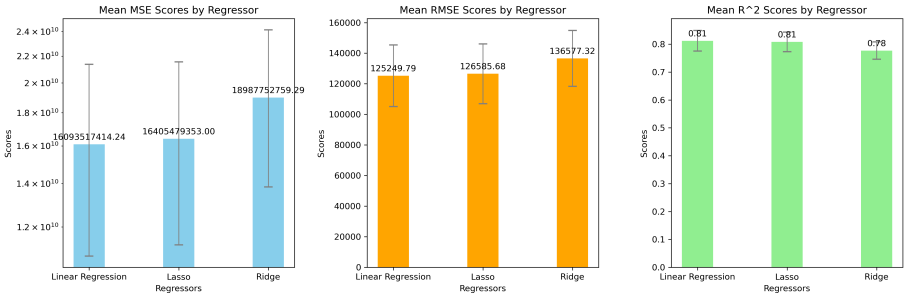

Figure 6: Bar charts showing various metrics across different regularizations

4.3.2 Nonlinear Models

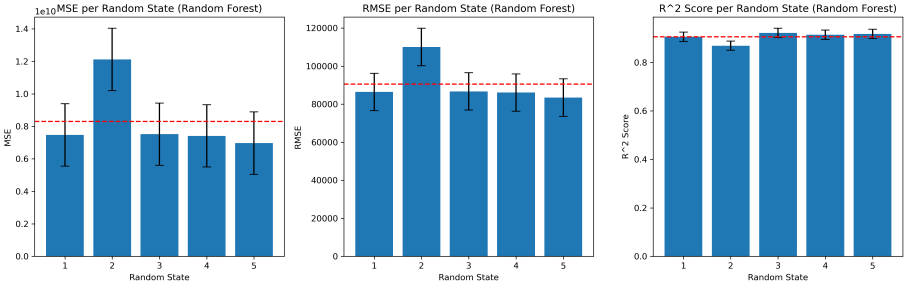

Figure 7: Random Forest. Bar charts showing performance across states.

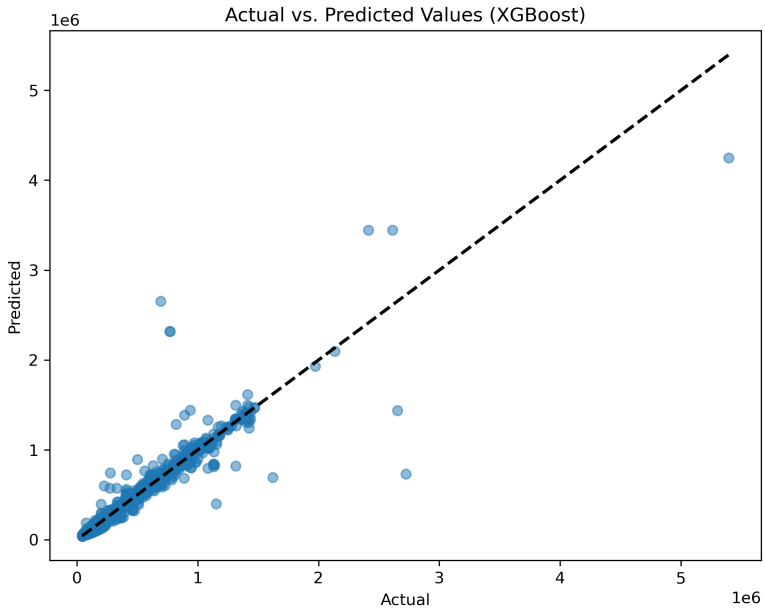

Figure 9: XGBoost. A scatter plot yielding similar results to RF with a slightly higher accuracy.

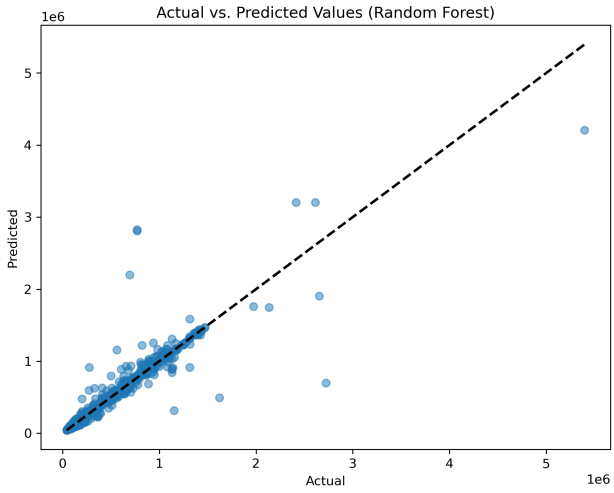

Figure 8: Random Forest. Scatter plot showing predicted vs actual values. Most points are densely clustered along the diagonal line, indicating relatively high accuracy with the exception of a few outliers.

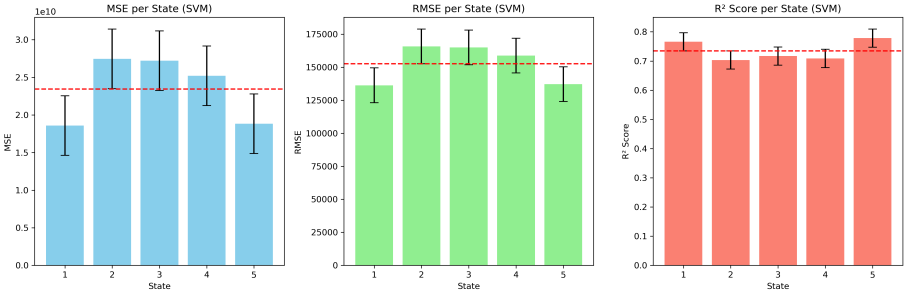

Figure 10: SVM. Bar charts showing performance across states.

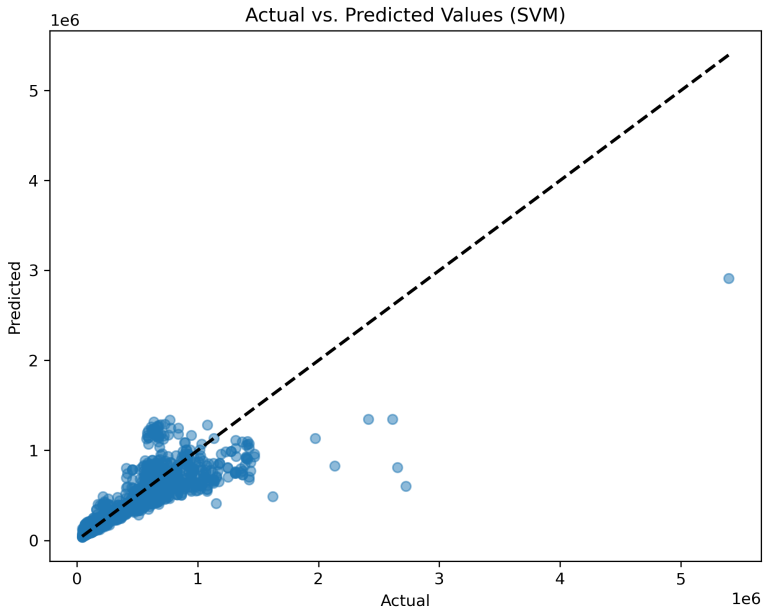

Figure 11: SVM. A scatter plot showing that, compared to the other algorithms, fewer points adhere to the predictive line, indicating lower accuracy.

4.4 Feature Importances

At least three different global feature importances were calculated.

Feature	Importance
dest_airport_code	0.7794
dest_country	0.1245
from_airport_code	0.0089

Table 5: Table showing the top three features contributing to prediction of emissions for Random Forest

Feature	Importance
dest_airport_code	0.2385
price	0.0584
dest_country	0.0266

Table 6: Table showing the top three features contributing to prediction of emissions for XGB

Feature	Importance
duration	1.3247
price	0.1714
aircraft_type	0.0385
departure_time	0.0165
dest_airport_code	0.0067
arrival_time	0.0012
dest_country	-0.0051
from_country	-0.0110
from_airport_code	-0.0124

Table 7: Using the RF pipeline, the table lists feature importances from highest to lowest.

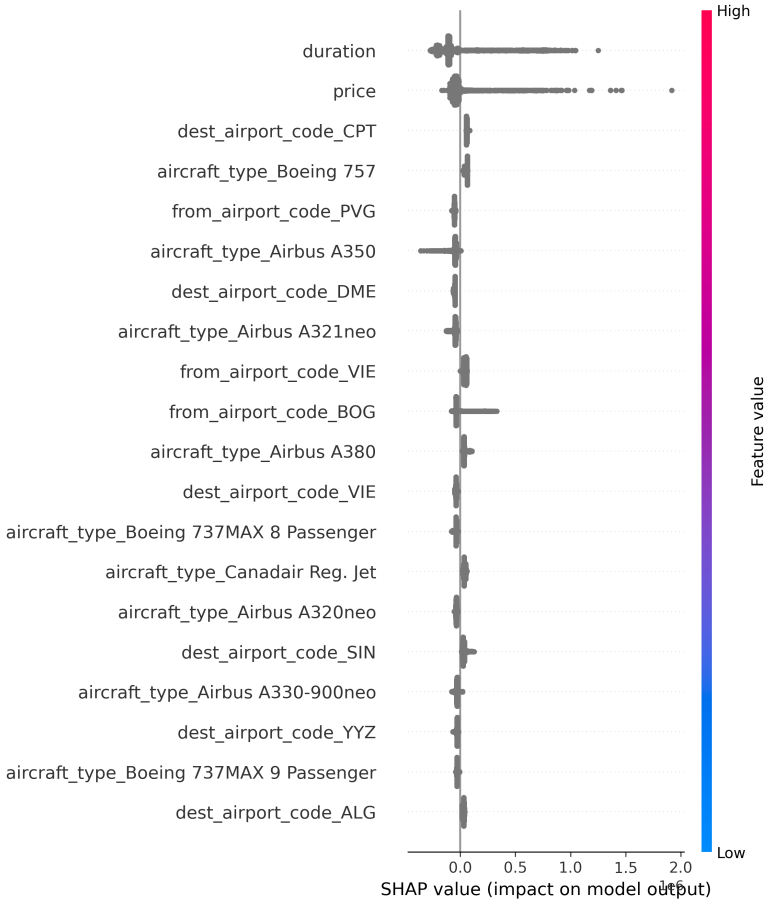

Figure 12: SHAP summary plot for XGB.

Aside from global feature importances, local feature importances were calculated using SHAP (see Figure 12). The XGB model was selected using the tuned hyperparameters. Duration and price were the most relevant features at the top of the plot, indicating that they are most influential in determining predictions for emissions. The clusters to the left of 0.0 for both these features suggest that, when at certain values, the features lead to a decrease in the predicted emissions relative to the model's base predictions. The thin spread across a wider range on the positive side of the X-axis suggests that in some cases, increases in duration and price can lead to an increase in predicted emissions. However, these cases might be less frequent (as indicated by the thinner spread) than those where these features have a small or negative impact.

The selection of the data points at index 0 for creating a force plot serves to display local feature importances (see Figure 13). There are three main features contributing positively to the prediction: price, from_airport_code_BOG, and duration. This suggests that when predicting emissions, longer and expensive flights increase the model's emission estimate for this datapoint.

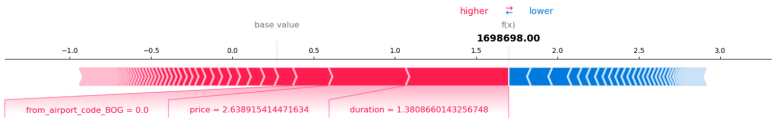

Figure 13: Force plot illustrating the three main features

4.5 Model Interpretations in Context

These model interpretations show how duration and price, as influential factors, can align intuitively with the nature of flight emissions. Longer flights typically have higher fuel consumption, which could lead to higher emissions, while the relationship with price could reflect flight route complexities, types of aircraft used, etc. The varied impact, shown through the spread in the SHAP summary plot, emphasizes how multifaceted emissions predictions can be.

5 Outlook

As mentioned in Section 4.2, a possible area of improvement is addressing the lack of standard deviation in XGBoost by experimenting with different ways of splitting data. Furthermore, the incorporation of more diverse data (such as specific aircraft fuel efficiency) could refine the accuracy of the model. New insights could also be garnered through feature engineering. Exploring alternative algorithms is also a point of improvement, such as using MLPRegressor. This was attempted during implementation of the project, but due to high complexity and extensive runtimes, refining this model would only be possible given additional time.

6 References

- Demir, A.S. (2022). Modeling and forecasting of CO2 emissions resulting from air transport with genetic algorithms: the United Kingdom case. *Theor Appl Climatol*, 150, 777–785. https://doi.org/10.1007/s00704-022-04203-4

- Polartech. (n.d.). Flight data with 1 million or more records. Kaggle. Retrieved from https://www.kaggle.com/datasets/ polartech/flight-data-with-1-million-or-more-records