EcoFly: What Factors Most Strongly Predict CO_2 Emission Performance?

Group 5: 31696, 28521, 29678

ABSTRACT

The environmental impact of the aviation industry is increasingly under scrutiny, but what are the factors which most strongly influence CO_2 emissions in the industry? This paper seeks, through explorative analyses, to determine which factors attached to a flight route—amongst these being the airline, aircraft, departure and arrival airports, price and beyond—are the strongest predictors of a route's CO_2 performance compared to the average for that route. We find through a logistic regression that airlines play a strong role in determining the probability of a route performing better than average, and through an ensemble of decision trees on a balanced outcome class dataset that aircraft types are the most important features in predicting CO_2 performance. We direct future research towards latent dimensionality investigation and feature engineering to extricate intra-predictor relationships.

CCS Concepts

• **Computing Methodologies→Machine Learning →Supervised Learning.** Leveraging supervised machine learning models, including linear and ensemble methods

• **Information Systems→Data Management Systems→Big Data Infrastructures.** Leveraging distributed systems to filter, aggregate and process data. Parallelization of models across the networks to facilitate handling of large data while leveraging fault-tolerance in the network.

• **Software and its Engineering→Software Creation and Management→Software Development.** DataFrame API and MapReduce in Spark facilitates data aggregation, filtering and transformation. Leveraging the Spark ecosystem allows more efficient data processing across nodes in the network.

Keywords

CO_2 Optimization, Environmental Impact, Aviation, Flight Routes, Apache Spark, MapReduce, Big Data, Machine Learning.

1. INTRODUCTION

In the dynamic landscape of aviation, the optimization of flight routes plays a crucial role for airlines and pilots alike. While fuel is the second largest cost incurred by airlines, accounting for 22% operational expenses [3] flight routes are often changed to optimize fuel consumption, which is reflected in the prices proposed to consumers. The objective of this study is to determine what factors most strongly predict emission performance. This has a strong policy implication both for the private aviation sector which can increase the optimization of its routes and subsequent emissions but also for the public sector where emission caps are proposed. Different airlines are investigated, aiming to uncover which of them are most susceptible to lowering their emissions and what covariates are to be accounted for this change. We use the *Global Flight Data* dataset, harvested by BarkingData, by scraping Google Flights and available on Kaggle [2]. In this report we first recount the literature on the topic, paying particular attention to the previous applications of distributed systems in this context. Second, we provide a broad overview of the dataset's characteristics through a descriptive analysis, including a social network of airports. We derive from this a community detection algorithm which provides preliminary insight into airport groupings based on emissions which correspond primarily to matching airlines. We use the results of this analysis to motivate the following section; predictive modelling. Different modelling methods are employed, aiming to answer a following research question: *what factors most strongly predict CO_2 emission performance?* Our approach, which leverages distributed computing and bagging for ensemble methods allows us to account for other factors impacting emissions size, such as aircraft type and duration of a flight. Our report finds that when producing a regularized logistic model predicting whether or not a flight will perform better than average in terms of emissions, the main predictors of a positive effect are airlines. We discuss at length the reasons we attribute to this as well as the limitations of this basic approach. We supplement this linear model with a more complex bagging method of decision trees on an under-sampled majority class dataset from which we extract the most important predictors. We find that aircraft types, price and the number of stops are the most important features. The report concludes by summarizing limitations with respect to data selection and sampling, our methods, and future directions for research.

2. LITERATURE REVIEW

The environmental impact of flight routes has become an area of increasing concern for customers, policymakers and airline managers. Notably, the air quality improvement resulting from the Covid-19 lockdowns [22] have increased scrutiny in this area and prompted the airline community to research new solutions to improve the situation without resorting to fight restrictions. Beyond environmental costs, a poorly optimized sky route is costly for passengers and airlines alike, further promoting research in this area. Before presenting our data and methodology, we proceed by providing an account of the literature on the matter thus far. We identify two relevant branches of literature: i) that pertaining to the small-scale predictive modelling for aviation industry pollution, and ii) leveraging distributed systems to handle large-scale environmental and aviation data. By exploring *which flight carrier is most likely to optimize its CO_2 emissions,* we aim to bridge these two branches of the literature, conjoining the efforts thus far made in predictive modelling specific to the aviation industry's emissions and scalable machine learning with large data.

2.1 Emissions Modelling

Various scholars have thus far attempted to use aerospace data to model a variety of outcomes including, CO_2 emissions, flight prices [18], delays [12] or even long-term environmental effects of air traffic-related pollution [14]. A variety of machine learning models are deployed in this area; in a study of CO_2 consumption per aircraft, [16] propose the use of clustering algorithms to extract flight characteristics which produce similar levels of CO_2 emissions amongst airline profiles. The study uses historical data by EUROCONTROL's Base of Aircraft Data (BADA) on the Climb-Cruise-Descent (CCD) cycle (excludes taxi before take-off and after landing) flights in the US airspace to explore different

clustering algorithms that will predict the fuel use of an aircraft at the different stages of the CCD cycle. The high variability in their findings serves to confirm that aircraft type, relative altitude on the same route, aircraft weight and management result in very different fuel efficiencies for the same route. Sheng, Marais and Landry (2015) propose similar findings but for the stratospheric fuel consumption of the aircraft, ie. above a certain altitude, which is argued to be more harmful than the rest of the aircraft's emissions. However, these does not provide us with information about which airlines are more likely to optimize these differences, and does propose any scalable solutions to extend these findings outside of the US-context. On the other hand, Vergnes *et al.* (2022) contend that in comparison to their unsupervised geo-spatial analysis of airplane routes worldwide, the current routes are already close to optimal with regards to using winds to the pilots' advantage and optimizing altitude for air pressure. However, the scope of this analysis is limited to direct flights. Furthermore, their proposed model suggests that a 5% improvement in route optimization is still possible for mid-haul flights in Europe specifically. Relatedly, it had been previously investigated by Alcabin *et al.* (2012) that there was room for at least a 3% emission improvement in vertical flight path specifically (when the aircraft is either moving up or down, not the cruising at high altitude) for U.S. flights. However, along with the other studies deploying machine learning to explain the intricacies of emissions with other airline-level variables, this analysis is limited to small datasets and does not leverage distributed systems.

2.2 Leveraging Distributed Systems

We identify two main ways in which the current literature explores the use of distributed systems in relevance to our goal of predicting which airlines are most likely to optimize their CO_2 emissions. The first leverages MapReduce and Spark to handle general environmental and climate-related data, but without a specified application to CO_2 emissions by aircrafts. The second uses these same infrastructures for the aviation industry but not specifically to CO_2 emissions. We summarize the progress in these two categories before outlining how our stances contributes to these advancements.

2.2.1 MapReduce, Spark and the Environment

In recent years, there has been a marked increase in the accessibility of environmental data, which are frequently characterized by complex temporal dimensions that not only enrich the data but also significantly expand its volume, making it ideally suited for leveraging through distributed systems [7]. In fact Xu et al. (2020) propose a hybrid framework for wind speed data forecasting across China, leveraging Apache Spark to parallelize large-scale operations. While the study was intended to promote wind-sourced energy, the applications extend far beyond, including more wide-spread meteorological forecasting [13] used by the aviation industry for route—and therefore CO_2 emission—optimization (Degaugue *et al.*, 2021). Hu *et al.* (2018) contend that even the current distributed system architectures like Apache Spark will grow insufficient for processing the ever-growing stack of climate data and propose ClimateSpark for big data storage and analytics. In addition to the functionalities proposed by Spark, ClimateSpark includes its own data model (ClimateRDD) which enables streamlined interaction with spatiotemporal indexing. Similarly, the Apache Open Climate Workbench [9] and the SciSpark big data framework [17] also seek to facilitate the analysis of climate and earth through the extension of Spark architectures specifically designed for environmental data. While these studies demonstrate the successful building and deployment of distributed systems to facilitate the analysis of climate data, these lack the specific applicability to the aviation industry which we seek to establish.

2.2.2 MapReduce, Spark and Aviation

The study of the aviation industry extends far beyond its environmental implications; it spans over predictive models for costs, and cancellations and delays. In fact, flight delay prediction is the most common reason for leveraging MapReduce and Spark in the literature, with numerous scholars using real-time data to dynamically predict delays [5, 24], and others focusing on historical weather data [4] or airline-level data [12] as delay predictors. Khotimah *et al.* (2023), through their review of the performance of these distributed algorithms reveal that by leveraging the full extent of the data through distributed systems, predictive models in the aviation industry offer a better performance compared to the small-scale algorithms proposed by other researchers in the field. Finally, it is noteworthy that research on flight safety and incident likelihood is also prominent in the literature [11, 20]. This sub-section of the literature exemplifies the importance of using distributed systems for scalable machine learning in the aviation sector, and further underlines that there is insufficient data combining these technologies for the purposes of emission optimization in this industry. Thus, we seek to combine the findings from the successful use of big data in aviation research with flight data containing environmental impact information for each airline to contribute to the literature.

3. METHODOLOGY

3.1 Data and Sampling Strategy

This analysis utilizes *Global Flight Data* dataset made publicly available on Kaggle by BarkingData. It includes information on 998, 865 flights, in particular their country of departure, country of arrival, code of the departure airport, code of the arrival airport, aircraft type, airline number, airline name, flight number, departure time, arrival time, duration of the flight, number of stops if flight is not direct, price at an economic fare, currency, actual CO_2 emissions per passenger in that specific flight, average CO_2 emissions per passenger on the route, percentage difference between the two, date of when the information was scanned. Investigated flights have departed between 30th April 2022 and 28th August 2022.

For this analysis, in order to increase comparability of flights, solely information on flights departing and arriving on biggest airports per continent, available in the dataset, was subsetted. Selected airports are: Beijing Capital International Airport (PEK), Paris Charles de Gaulle Airport (CDG), Toronto Pearson International Airport (YYK), Sao Paulo/Guarulhos International Airport (GRU), Sydney Airport (SYD) and Cairo International Airport (CAI).

Moreover, intended for the modelling segment, additional dataset with further manipulations was created. Solely variables of interest for modelling were kept, i.e. code of the departure airport, code of the arrival airport, aircraft type, airline name, duration of the flight, time of departure, time of arrival, number of stops, price and

CO_2 percentage difference between the actual emission per person and the average emission per person for this route. Subsequently, times of departures and arrivals have been separated into 4 categories: 'Morning', 'Afternoon', 'Evening' and 'Night' as those are often associated with different levels of airport congestion. Moreover, categorical data has been one-hot encoded.

3.2 Methods

The study uses a machine learning approach to predict whether a flight has an improved, i.e. better than average for that route, CO_2 emission. Firstly, regularized linear regression model is fitted, allowing to explore the continuous nature of the outcome variable, and predicting the % of CO_2 improvement/deterioration. Followingly, feature coefficients are explored to investigate each predictor's contribution and allow identification of factors, specifically individual airlines that tend to have better than average CO_2 emissions. As the study optimizes model performance over specific numeric information on the extent of CO_2 emissions improvement/deterioration, the outcome variable is then recoded to solely binary representation of whether the CO_2 emission was improved or not. To explore this classification problem, regularized logistic regression is first fitted, and its' predictors' coefficients investigated for specific coefficient contribution. To improve our model further, we address class imbalances by under-sampling the majority and find an optimal ratio that maximizes recall. Finally, the optimally balanced training data is then used to develop an ensemble of decision trees with bootstrapped aggregating. This involves training multiple models on different random subsets of the training data independently, and applying a voting strategy that computes the consensus across all classifiers. Feature importance is extracted by computing importance scores across all 10 models and ranking the top features.

4. NUMERICAL RESULTS

4.1 Descriptive Statistics

Study employs MapReduce techniques to explore the data descriptively.

4.1.1. Airports Congestion

Figure 1: Top 10 Departure Hubs

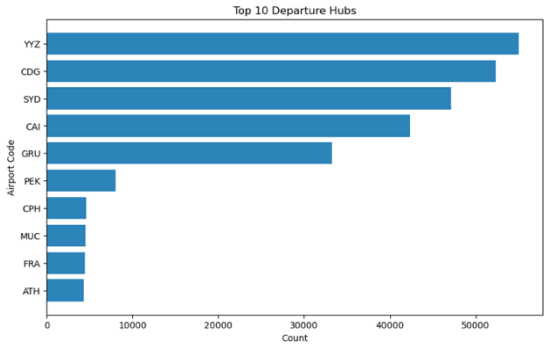


Figure 2: Top 10 Arrival Hubs

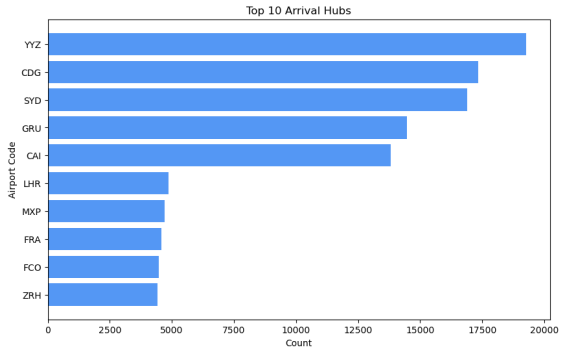


Figure 1 and *Figure 2* present the busiest hubs within our dataset, regarding number of departing and arriving flights. Airport congestion plays a big role in an aircraft's CO_2 emission. That is due to taxi time, i.e. how much time an aircraft spends circulating around the airport, waiting in a queue to depart, as well as circulating above the airport before being able to land. While longer taxi times are often caused by weather conditions, they also reflect airport congestions (Eurocontrol, 2023).

4.1.2. Number of Stops

Table 1: Distribution of number of stops

Stops	0 (Direct)	1	2	3	4	5	6
Count	5426	130, 204	156, 597	13, 116	647	34	1

A diverse distribution of number of stops, as visualized in *Table 1*, allows a broader exploration on whether direct flights are always the most sustainable, as often claimed in studies (Debbage and Debbagge, 2019; Rhodes, 2022). Most flights have 1 or 2 stops, which shifts the focus of our work to include these as well as direct flights. We control for this in our model instead of removing non-direct flights; the intuition is here that a direct route is not necessarily more fuel efficient than a layover. We do not, however, aim to account for other elements like time spent traveling or other factors which may render direct routes overall preferable; the focus is solely on emissions.

4.1.3. Most Popular Routes

Table 2: Top 10 most popular routes

Departure Airport	Arrival Airport	Number of flights
Paris Charles de Gaulle	Zurich Airport	1159
Paris Charles de Gaulle	Rome Fumicino	1158
Paris Charles de Gaulle	Amsterdam Schiphol	1157
Paris Charles de Gaulle	Copenhagen Airport	1152
Paris Charles de Gaulle	Milan Malpensa	1138
Paris Charles de Gaulle	London Heathrow	1129
Toronto Pearson	G. Bush Airport (Houston)	1127
Munich Airport	Paris Charles de Gaulle	1118
Toronto Pearson	Milan Malpensa	1117
Paris Charles de Gaulle	Athens International Airport	1111

Table 2 presents the top 10 most popular routes present in the dataset. It shows that even though one airport per continent has been selected, the data is highly biased in the European side. That,

however, is not surprising as out of the selected airports, Paris Charles De Gaulle is the busiest one.

4.1.4. *Emissions by Airline Combinations*

Table 3: Top 10 Emission Contributors by Airline Combinations

Airline Names	Average Emissions/ in millions
Lufthansa - Singapore	10.5
China Eastern – Jet Blue – American	9.8
China Eastern – Delta – United	9.2
LATAM – ANA	9.0
American – ANA – JAL	9.0
British Airways – Finnair – JAL	9.0
Lufthansa – British Airways	9.0
LATAM – Alaska – United	8.9
Qantas – Avianca – LATAM	8.5
LATAM – American – JAL	8.5

Table 3 answers which airline groups contribute most to emissions in our dataset. By grouping routes by their unique airline combinations, we find that Lufthansa – Singapore routes have the highest average emissions of 10.5 million kg/person. This could be due to a myriad of factors which we will explore further in subsequent sections, but what is noteworthy is that top emitting airline combinations have varying number of stops, thus setting the stage for further analysis on how number of stops affects emissions.

4.1.5. *Emissions by Flight Stops*

Table 4

Stops	0 (Direct)	1	2	3	4	5	6
Average Emissions (in millions)	0.31	0.82	1.43	1.95	1.79	1.37	2.32

Table 4 gleans insights into how average emissions vary by number of stops. Unsurprisingly, there is a general trend that as the number of stops increases, the average emissions increase as well. Direct flights have the lowest emissions, since they are most probably the flights with the shortest duration and distance.

4.1.6. *Correlations between Factors*

Table 5

	CO2 Emissions	Price	Duration	Stops
CO2 Emissions	1.0	0.68	0.51	0.36
Price	0.68	1.0	0.40	0.31
Duration	0.51	0.40	1.0	0.52
Stops	0.36	0.31	0.52	1.0

Assessing correlations between key features we feed into our training model is crucial in ascertaining the suitability of our machine learning approach, and determine if there are issues with multicollinearity. These insights also help in understanding the interplay between different factors influencing CO2 emissions.

Table 5 shows correlations between some of the factors – we observe modest positive correlations between the continuous features. Specifically, more expensive flights and longer flights have higher CO2 emissions. The presence of multicollinearity can lead to issues in the linear model, and can destabilize the model estimates. Therefore, results must be interpreted with caution.

4.1.7. *A Route Network Model*

Figure 3: Network Graph of Airports sized by in-degree Centrality and Weighted by Emissions per Minute

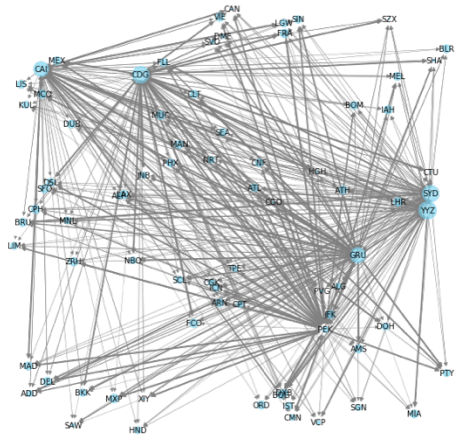


Figure 3 illustrates a network of the airports in the dataset, with their nodes sized by in-degree centrality, and the edges between them weighted by the emissions per minute of that journey. Some hubs are apparent in the network, such as YYZ (Toronto), GRU (São Paulo) and SYD (Sydney) or CDG (Paris) in Europe. These airports, relative to the others, have some of the highest number of in-coming flights. What they have in common beyond this, however, is that they generally are part of flights whose emissions per minute are also higher relative to other flights. Considering that these flights are national hubs as much as they are international ones, they engage in both domestic, short distance flights and long-haul flights. This means that the current emissions per minute, which are generally higher than other airports' for the routes these airports belong to, involve other factors than simply distance (and therefore air pressure, winds and other factors listed in the literature as contributors to emissions). Our subsequent analysis identifies the main predictors of CO_2 performance and therefore the factors which contribute to these differing emission performances amongst routes. Below is a preliminary community assignment of the airport to motivate further research into the effects of airlines and aircraft types on the emissions of a route.

Figure 4: Network of Airports by Greedy Modularity Communities Algorithm, weighted by CO_2 emissions per minute per route.

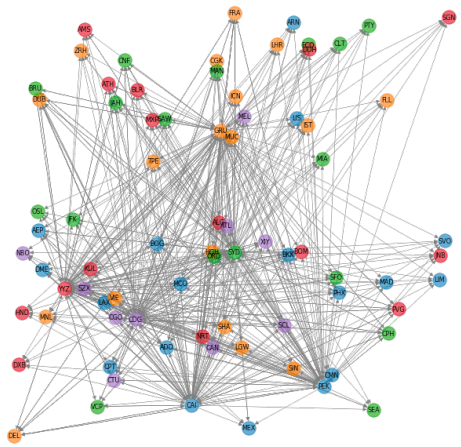


Figure 4 presents the detected communities from the nodes in *Figure 3* based on the Greedy Modularity Communities algorithm, which considers the CO_2-weighted edges between nodes. This algorithm starts with every node as its own community and iteratively merges them to maximize modularity gain (strength of division of the network) where connections between members of the same community are maximized, but sparse between members of different communities [15]. This algorithm is particularly useful in this case because it accounts for edge weights, which in this case represent emissions per minute. The algorithm thus forms communities based on the emissions between airports, with airports in the same community exhibiting similar emission patterns in the routes they are involved in. This is for both direct and non-direct flights.

Figure 4 illustrates a number of communities, amongst our selection of airports, which are constructed in part based on the weighting of the emissions per minute on the routes between them. This yields five primary communities. The first group which carries particular saliency is indicated by the purple nodes. This community regroups the major Chinese airports and one overseas connection in each continent. As this is the product of a grouping largely based on similarities between airport route participation and the attached CO_2 emissions per minute, this is a good descriptive indication that airlines which operate in China have similar emission patterns on their routes. This could be due to the national airline serving these routes using the same types of aircrafts or fuel optimization techniques for its routes which involve these airports. Secondly, the blue nodes' community is largely made up of Latin American airports and their connections to other continents. This is a group of airports primarily served by the same airlines in the *OneWorld* alliance. While airline alliances do not always use the same aircrafts for the same routes, they benefit from other factors which could contribute to their grouping in a network community based on their emissions per minute on a given route, such as pilot training, joint maintenance and repairs costs as well as fuel choice and optimization strategy. It is thus not surprising, though no causal inference can be made from this network, to see that airports' community groupings are to some degree aligned with regions and airlines. As for the other communities (orange, red and green), we can see similar patterns within the orange group which regroups many hubs of the Star Alliance fleet, or the red community regrouping a mix of Star Alliance and SkyTeam airlines. The green community represents most trans-Atlantic routes operated by the main North American and European carriers. While some of these communities are looser, it is noteworthy that some degree of grouping along these characteristics exist. This further motivates our modelling in the next section which (see Table 5) identifies that airlines are indeed the strongest predictor of whether a flight on a route will emit less than average. The groupings formed based on the greedy algorithm, weighted for emissions, thus reflect some of our later findings with regards to the importance of airlines and their corresponding alliances which tend to operate regionally, and produce similar effects on emissions. Furthermore, the groupings by airlines and alliances provides a good starting point for further research (see section 5) for flight recommendations.

4.2 MACHINE LEARNING RESULTS

4.2.1. Linear Models

Firstly, the following relationship is explored:

CO_2 % difference ~ Airline + Aircraft + Price + Duration of the flight + Airport of Departure + Airport of Arrival + Time of Departure + Time of Arrival + Number of Stops

CO_2 % difference implies the difference between the actual CO_2 emission of the specific flight and the average one for the route. Therefore, it allows delving into the issue of improving flight's CO_2 emission.

Having divided data into train and test set in 80-20 proportion, we fit a regularized linear regression, aiming to predict *to what extent CO_2 in the investigated flight is better or worse than the average one for the route*.

Table 6: Linear Regression Model Evaluation

Evaluation Metric	Value
Mean Absolute Error	44.74
Root Mean Squared Error	73.45
R-squared	0.30

Table 6 presents performance of the fitted Linear Regression model. All three values are rather low, suggesting the model should be improved. We prioritize the maximization of model's performance over knowledge of the exact extent to which CO_2 emissions are improved/deteriorated; the outcome variable is recoded to binary (indicates whether emissions were better than the average for the route or not). This enables regularized Logistic Regression fitting.

Table 7: Logistic Regression Model Evaluation

Evaluation Metric	Value
Area under ROC	0.86
Area under PR	0.64

The AUC-ROC value indicates a good overall performance, with ranking a randomly chosen positive instance higher than a randomly chosen negative instance about 86% of the time. AUC-PR of 64%, however, indicates the model has moderate performance in terms of the trade-off between precision and recall, likely stemming from class imbalance. For this reason, we develop further models where we address this issue.

Table 8: Predictors with 8 strongest positive coefficients

Predictor	Coefficient
Zambia Airways	4.169
T Way Air	4.166
Indonesia AirAsia	3.937
FlyEgypt	3.890
Smartwings	3.824
Start Flyer	3.395
Air Botswana	3.365
Hong Kong Express	3.108

Table 8 reveals the predictors with the 8 strongest positive coefficients. Here, positive coefficients represent a positive relation between the airlines and our outcome variable; a change in CO_2 emission compared to the average for the route (in this case a positive value is a decrease compared to the average). Interestingly, these predictors are all airlines, suggesting that these particular flight carriers are most likely to have improved CO_2 emissions. We concede that the airline variable indirectly captures information about aircraft type and price, but through regularization we mitigate the effect of this. It is also important to note that there is a significant outcome class imbalance in the data, with the majority class being a lack of improvement in comparison to the benchmark average. This linear model does not mitigate for this effect; we seek to achieve a comparatively better performance on our next model.

Table 9: Predictors with 8 strongest negative coefficients

Predictor	Coefficient
Belavia Air	-2.969
Jeju Air	-2.876
Destination Airport: Sheremetyevo Moscow Airport	-2.855
Allegiant Air	-2.770
Embraer ERJ 135 145	-2.683
Airbus A340	-2.588
TUI Airways	-2.436
Cayman Airways	-2.408

Table 9 identifies the 8 predictors with the strongest negative coefficients; the 8 predictors with the highest negative association to the outcome variable. These are the predictors for which the change in emissions compared to the average are negative (the actual emissions are higher than average). This list contains 5 airlines, 2 aircraft-types and an airport, suggesting all of those might impact CO_2 emissions negatively. Exploring the airlines, most of those on the list tend to use older aircrafts, prioritizing fuel efficiency and therefore CO_2 emission control less well than newer models. This could also be reflected in the aircraft features. Finally, Sheremetyevo (Moscow) Airport is amongst the most polluting airports due to its high average taxi times [21].

4.2.2. Targeted Model Improvements

Addressing Class Imbalances: Under-sampling Majority Class

Table 10: Binary Class Imbalance

Emission Improvement	Count
Improved (Class 1)	71,096
Not improved (Class 0)	225,125

Table 10 reveals clear outcome class imbalance in our dataset – there are over 3 times as many routes that have worse emission levels than average for a given route, which biases our estimates. Specifically, the models trained on unbalanced data is biased towards the majority class of routes with worse-than-average emissions, and will therefore have difficulty learning minority class patterns. This inevitably leads to poor out of sample performance. To address this, we under-sample the majority class. Given the substantive size of the data set and the exploratory nature of our work this approach is preferable to over-sampling of the minority class which duplicates observations or other K-Nearest Neighbors approaches to class balancing, as these result in synthetic data creation and in practice amplify existing biases in the data. In doing so, we limit the introduction of noise in our data and contribute to preventing over-fitting induced by synthetic data creation, as well as computational expenses introduced by methods like SMOTE on such a large dataset.

To resample, we first split the data using stratified sampling, ensuring the fraction of the minority in the training and test size are similar. The baseline model is a Random Classifier to compare against the against the resampled data (see *Figure 5* below).

Figure 5: Random Classifier Model Performance

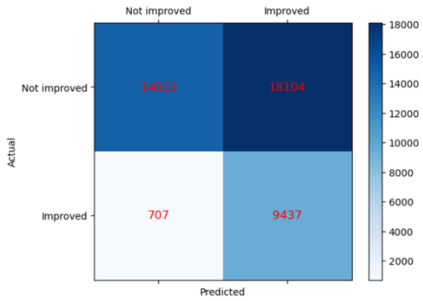
Figure 5 illustrates that the model has a high accuracy, mainly derived from the high number of True Negatives. Recall, however, is incredibly low, suggesting the model cannot correctly predict routes with improved emissions, which is what we are more interested in. Using an under-sampling ratio of 1:1, meaning classes are equally-balanced, we fit Random Forest Classifier and compare against the baseline as shown in *Table 8*.

Table 11: Baseline Model vs. Balanced Model Evaluation Metrics

Evaluation Metric	Baseline Model	Balanced Model
Area under ROC	0.81	0.82
Area under PR	0.57	0.56
Recall	0.00	0.93

Table 11 above indicates that under-sampling significantly improved recall of the model to 93%, meaning the model predicts true positives much better. *Figure 6* below illustrates the prediction success of the model.

Figure 6: Confusion Matrix for Balanced Model



Ensemble Learning with Bagging

With the balanced train dataset, we perform ensemble learning with decision tree classifiers. We create a bootstrap sample from the training features with replacement, where each instance has an equal chance of selection. There are 56929 instances of class 1, and 86748 instances of class 0. Using a decision tree as our weak learner, we build an ensemble with 10 iterations and build a pipeline with the models as stages. Finally, we aggregate predictions (through a user-defined function to compute a consensus) across all the classifiers.

Table 12: Ensemble Model Performance Metrics

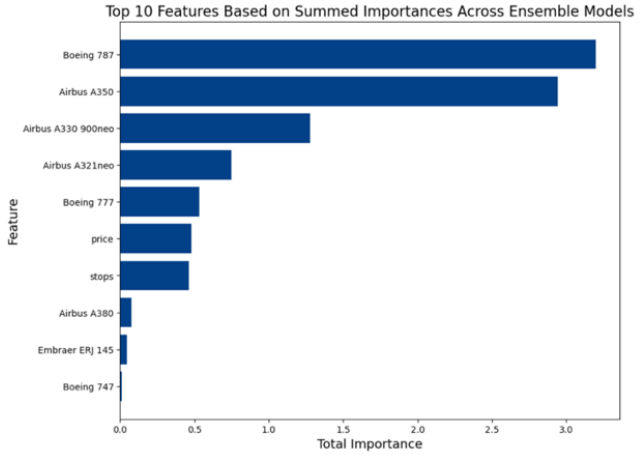
Evaluation Metric	Value
AUC	0.75
Accuracy	0.74
Precision	0.81
Recall	0.74
F1	0.76

Table 12 illustrates the overall good performance of the final model in distinguishing between the classes with a decision boundary at 0.5. The ensemble learner is a clear improvement from the unbalanced model with a good balance between recall and precision.

Feature importance from Ensemble

Extracting feature importance across the ensemble models is a useful way to understand which factors drive CO_2 emission performance. Feature importance was extracted from each model, and importance was summed up for each predictor and sorted in descending order. *Figure 7* below summarizes the top 10 features:

Figure 7: Ensemble Model Feature Importance



Unsurprisingly, aircraft carriers are significant features in emission improvement. Boeing 787 is the most important feature, which resonates with its reputation of being a super-efficient family of commercial airplanes, boasting 25% reduction in fuel consumption compared to earlier models. While airlines were identified as the most important features in the linear model, we now see that the main features in our balanced model are aircrafts. While the descriptive analysis and our first model identified airlines as being crucial to a decrease in CO_2 it is plausible that these two features are co-linearly involved. This suggests that there may be overlapping dimensions of these two variables which we cannot account for with our current feature selection. Importantly, price and the number of stops are still important predictors of CO_2 performance, though this effect seems to be lesser than what the literature originally proposed. We discuss at lengths the limitations which of these findings in the next section.

5. LIMITATIONS

Our modelling strategies are subject to multiple limitations which caveat our findings. Firstly, we must re-iterate that the sampling strategy includes only one airport per continent; thus our approach is limited in scope to this distribution of airports and their operating airlines. This means that the analysis leaves out a number of carriers, destinations, aircraft types and their associated co-variates. Nonetheless, our findings provide a benchmark for future research in this field to encompass a larger sample or even real-time data in the aviation industry.

Second, though we control for flight duration and thus measure emissions in accordance to it, the duration of the flight does not include taxi times or other time an aircraft spends on the ground with its engine running (luggage haul, and un-haul, technical checks). We thus control for it but acknowledge that there are limitations attached to its use as a predictor. Nonetheless, it does not return in the list of the most influential features; these effects are thus minimal on model performance and its interpretative value.

Thirdly, we include both airlines and aircraft types as predictors in our models. Though Random Forests are robust to multi-collinearity, and our logistic model is regularized, this does not directly address the multi-collinearity possible between airlines and aircraft types. We assume these two predictors to have sufficient non-overlapping dimensions to be relevant, separate predictors in this analysis, but this remains a limitation in the interpretability of our results and their general robustness. Further research should address this issue further, through feature engineering, perhaps by creating hybrid variables which capture information on the airline and the aircraft simultaneously. An exploratory analysis on this matter is warranted for future research.

Fourth, it is noteworthy that we use a general 0.5 threshold for our model evaluation due to the exploratory nature of this work. It is, nonetheless, warranted to further research the effects of different decision thresholds on model performance and investigate whether features' importance varies with different thresholds.

Fifth, our analysis focuses on web-scraped flight data, which is subjected to change over time, particularly in terms of price as fuel prices and the economy change. Differences in these factors could plausibly result in different strengths of effect on CO_2 emission performance. An improvement to this limitation would constitute working with real-time data as opposed to historic data or future data as in the case of web-scraping. Nonetheless, we expect the price estimates to have relatively minor fluctuations and thus relatively minor effects on model performance and feature importance. It is also noteworthy that because scraped flight data is future data in nature, we are unable to account for other delays to emergency landings which result in different emission levels. We are only able to, at this time, distinguish a flight's planned performance from the average planned performance of all flights on that given route. This is a substantive limitation of our classification algorithm and applicability of findings, though we primarily seek to provide a benchmark for targeting different factors when attempting to optimize emission performance.

Finally, we seek to improve on our linear model using a Random Forest with Bagging. While bagging enables us to take advantage of distributed systems, further research is needed in deploying Boosting methods to gain a more comprehensive overview of the relative appropriateness of different methods when working with aviation data. Nonetheless, we leverage through Bagging both the breadth of data available and the efficiency of distributed systems, which helps us bridge the gaps identified in the literature, and maintain scalability in our solution.

6. CONCLUSION

In conclusion, this paper sought to answer *what factors are the strongest predictors of* CO_2 *performance*. We find, through the deployment of a more naïve, preliminary logistic model that

airlines are the primary determinant of improved CO_2 performance. This aligns with our descriptive analysis which reveals community groupings of routes which are aligned, through CO_2 emission similarities with airlines and alliances. However, with a more sophisticated ensemble model which we deploy on a balanced dataset that aircraft categories, as well as price and the number of stops are the most important predictors of CO_2 performance. Considering the literature and the general intuition that CO_2 performance is primarily derived from fuel efficiency in the first place, these findings make sense. The number of stops is also frequently mentioned in the literature as a factor affecting CO_2 performance, and it is thus unsurprising to find it in the most important features here. With regards to price, we note that price is importantly linked with the efficiency capabilities of an airline, and the associated aircraft. It is also noteworthy that while with two different approaches we obtain different features of importance, this may be due to overlapping dimensions of our features which are not adequately captured by the model. We stress that further research into how to most effectively engineer features is crucial here to further extricate the relationship between CO_2 performance and the factors which drive it, without the confounding potential of latent dimensions to the data. Overall, we find that our results are largely and plausibly in line with the relationships suggested in the literature, and propose to further this research through the expansion of distributed systems' use in the context of aviation research.

7. STATEMENTS

Statement of individual contributions

All members of the group contributed equally to theoretical and technical components.

Statement on the use of AI

The project was conducted using only course materials and independent research cited where appropriate in the code.

8. REFERENCES

[1] Monica Alcabin, Robert Schwab, Susan Cheng, Kwok-On Tong, and Charlie Soncrant. 2012. Measuring Vertical Flight Path Efficiency in the National Airspace System. In *9th AIAA Aviation Technology, Integration, and Operations Conference (ATIO)*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2009-6959

[2] BarkingData. 2022. Flight Data with 1 Million or More Records. Retrieved April 26, 2024 from https://www.kaggle.com/datasets/polartech/flight-data-with-1-million-or-more-records

[3] Brian Beers. 2023. Which Major Expenses Affect Airline Companies? *Investopedia*. Retrieved April 29, 2024 from https://www.investopedia.com/ask/answers/040715/what-are-major-expenses-affect-companies-airline-industry.asp

[4] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. 2016. Using Scalable Data Mining for Predicting Flight Delays. *ACM Trans. Intell. Syst. Technol.* 8, 1 (July 2016), 5:1-5:20. https://doi.org/10.1145/2888402

[5] Gerrit Burmester, Hui Ma, Dietrich Steinmetz, and Sven Hartmannn. 2018. Big Data and Data Analytics in Aviation. In *Advances in Aeronautical Informatics: Technologies Towards Flight 4.0*, Umut Durak, Jürgen Becker, Sven Hartmann and Nikolaos S. Voros (eds.). Springer International Publishing, Cham, 55–65. https://doi.org/10.1007/978-3-319-75058-3_5

[6] Sarah Degaugue, Jean-Baptiste Gotteland, and Nicolas Durand. 2021. Forteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2021) Learning Uncertainty Parameters for Tactical Conflict Resolution. In *14th USA-Europe Air Traffic Management Seminar*, September 2021. New Orleans, United States. Retrieved April 18, 2024 from https://hal.science/hal-03484004

[7] Daniel Q. Duffy, John L. Schnase, John H. Thompson, Shawn M. Freeman, and Thomas L. Clune. 2012. Preliminary Evaluation of MapReduce for High-Performance Climate Data Analysis. April 16, 2012. Pacific Grove, CA. Retrieved April 18, 2024 from https://ntrs.nasa.gov/citations/20120009187

[8] Fei Hu, Chaowei Yang, John L. Schnase, Daniel Q. Duffy, Mengchao Xu, Michael K. Bowen, Tsengdar Lee, and Weiwei Song. 2018. ClimateSpark: An in-memory distributed computing framework for big climate data analytics. *Computers & Geosciences* 115, (June 2018), 154–166. https://doi.org/10.1016/j.cageo.2018.03.011

[9] M. Joyce, P. Ramirez, M. Boustani, C. A. Mattmann, S. Khudikyan, L. J. McGibbney, and K. D. Whitehall. 2014. Apache Open Climate Workbench: Building Open Source Climate Science Tools and Community at the Apache Software Foundation. 2014, (December 2014), IN22A-06.

[10] Husnul Khotimah, Baiq Wilda Al Aluf, Muhammad Ari Rifqi, Ari Hernawan, and Gibran Satya Nugraha. 2023. Performance Analysis of the Distributed Support Vector Machine Algorithm Using Spark for Predicting Flight Delays. *E3S Web of Conf.* 465, (2023), 02037. https://doi.org/10.1051/e3sconf/202346502037

[11] S. Koteeswaran, N. Malarvizhi, E. Kannan, S. Sasikala, and S. Geetha. 2019. Data mining application on aviation accident data for predicting topmost causes for accidents. *Cluster Comput* 22, 5 (September 2019), 11379–11399. https://doi.org/10.1007/s10586-017-1394-2

[12] Khajjayam Sumanth Kumar, Kolipakula Karthik, Jeppiaar Nagar, and Rajiv Gandhi Salai. 2021. AIRLINE DATA ANALYSIS USING SPARK TECHNOLOGIES. (2021).

[13] Cosimo Magazzino, Marco Mele, and Nicolas Schneider. 2021. A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. *Renewable Energy* 167, (April 2021), 99–115. https://doi.org/10.1016/j.renene.2020.11.050

[14] Sigrun Matthes, Volker Grewe, Katrin Dahlmann, Christine Frömming, Emma Irvine, Ling Lim, Florian Linke, Benjamin Lührs, Bethan Owen, Keith Shine, Stavros Stromatas, Hiroshi Yamashita, and Feijia Yin. 2017. A Concept for Multi-Criteria Environmental Assessment of Aircraft Trajectories. *Aerospace* 4, 3 (September 2017), 42. https://doi.org/10.3390/aerospace4030042

[15] M. E. J. Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 23 (June 2006), 8577–8582. https://doi.org/10.1073/pnas.0601602103

[16] Ioanna Pagoni and Voula Psaraki-Kalouptsidi. 2017. Calculation of aircraft fuel consumption and CO2 emissions based on path profile estimation by clustering and registration. *Transportation Research Part D: Transport and Environment* 54, (July 2017), 172–190. https://doi.org/10.1016/j.trd.2017.05.006

[17] Rahul Palamuttam, Renato Marroquín Mogrovejo, Chris Mattmann, Brian Wilson, Kim Whitehall, Rishi Verma, Lewis McGibbney, and Paul Ramirez. 2015. SciSpark: Applying in-memory distributed computing to weather event detection and tracking. In *2015 IEEE International Conference on Big Data (Big Data)*, October 2015. 2020–2026. https://doi.org/10.1109/BigData.2015.7363983

[18] Ankita Panigrahi, Rakesh Sharma, Sujata Chakravarty, Bijay K Paikaray, and Harshvardhan Bhoyar. 2022. Flight Price Prediction Using Machine Learning. (2022).

[19] Hang Sheng, Karen Marais, and Steven Landry. 2015. Assessment of stratospheric fuel burn by civil commercial aviation. *Transportation Research Part D: Transport and Environment* 34, (January 2015), 1–15. https://doi.org/10.1016/j.trd.2014.10.008

[20] Donghui Shi, Shuai Cao, Jozef Zurada, and Jian Guan. 2022. *An Innovative Approach to Modeling Aviation Safety Incidents*. Retrieved April 18, 2024 from http://hdl.handle.net/10125/79482

[21] Shandelle Steadman and Sam Pickard. 2024. Airports, air pollution and climate change.

[22] Junzi Sun, Luis Basora, Xavier Olive, Martin Strohmeier, Matthias Schafer, Ivan Martinovic, and Vincent Lenders. 2022. OpenSky Report 2022: Evaluating Aviation Emissions Using Crowdsourced Open Flight Data. In *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, September 18, 2022. IEEE, Portsmouth, VA, USA, 1–8. https://doi.org/10.1109/DASC55683.2022.9925852

[23] Florent Vergnes, Judicaël Bedouet, Xavier Olive, and Junzi Sun. 2022. Environmental Impact Optimisation of Flight Plans in a Fixed and Free Route network. In *ICRAT 2022*, June 2022. Tampa, United States. Retrieved April 17, 2024 from https://hal.science/hal-03920682

[24] Gerasimos Vonitsanos, Theodor Panagiotakopoulos, Andreas Kanavos, and Athanasios Tsakalidis. 2021. Forecasting Air Flight Delays and Enabling Smart Airport Services in Apache Spark. In *Artificial Intelligence Applications and Innovations. AIAI 2021 IFIP WG 12.5*

International Workshops, 2021. Springer International Publishing, Cham, 407–417. https://doi.org/10.1007/978-3-030-79157-5_33

[25] Yinan Xu, Hui Liu, and Zhihao Long. 2020. A distributed computing framework for wind speed big data forecasting on Apache Spark. *Sustainable Energy Technologies and Assessments* 37, (February 2020), 100582. https://doi.org/10.1016/j.seta.2019.100582