INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Diogo Longo Polo

PREDICTIVE MODELING OF EN-ROUTE OPERATIONAL PERFORMANCE IN THE BRAZILIAN AIRSPACE

Final Paper
2025

Course of Civil-Aeronautics Engineering

CDU 656.7

Diogo Longo Polo

PREDICTIVE MODELING OF EN-ROUTE OPERATIONAL PERFORMANCE IN THE BRAZILIAN AIRSPACE

Advisor

Profa.Dra.Mayara Condé Rocha Murça (ITA)

CIVIL-AERONAUTICS ENGINEERING

São José dos Campos
Instituto Tecnológico de Aeronáutica

2025

Cataloging-in Publication Data
Documentation and Information Division

Polo, Diogo Longo
Predictive Modeling of En-Route Operational Performance in the Brazilian Airspace / Diogo Longo Polo.
São José dos Campos, 2025.
54f.

Final paper (Undergraduation study) – Course of Civil-Aeronautics Engineering– Instituto Tecnológico de Aeronáutica, 2025. Advisor: Profa.Dra.Mayara Condé Rocha Murça.

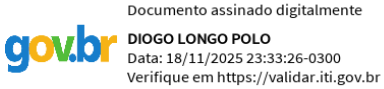
1. Controle do tráfego aéreo. 2. Planejamento estratégico. 3. Combustíveis. 4. Arvores de decisões. 5. Aprendizagem (inteligência artificial). 6. Segurança operacional. 7. Transportes. I. Instituto Tecnológico de Aeronáutica. II. Title.

BIBLIOGRAPHIC REFERENCE

POLO, Diogo Longo. **Predictive Modeling of En-Route Operational Performance in the Brazilian Airspace**. 2025. 54f. Final paper (Undergraduation study) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSION OF RIGHTS

AUTHOR'S NAME: Diogo Longo Polo
PUBLICATION TITLE: Predictive Modeling of En-Route Operational Performance in the Brazilian Airspace.
PUBLICATION KIND/YEAR: Final paper (Undergraduation study) / 2025



Diogo Longo Polo
Rua H8A, Ap. 127
12.228-460 – São José dos Campos–SP

PREDICTIVE MODELING OF EN-ROUTE OPERATIONAL PERFORMANCE IN THE BRAZILIAN AIRSPACE

This publication was accepted like Final Work of Undergraduation Study

Documento assinado digitalmente
gov.br
DIOGO LONGO POLO
Data: 18/11/2025 23:31:29-0300
Verifique em https://validar.iti.gov.br

Diogo Longo Polo

Author

Documento assinado digitalmente
gov.br
MAYARA CONDE ROCHA MURCA
Data: 19/11/2025 14:22:47-0300
Verifique em https://validar.iti.gov.br

Profa.Dra.Mayara Condé Rocha Murça (ITA)

Advisor

São José dos Campos: NOVEMBER 18, 2025.

To my grandfather Wilson Longo, who always believed in me and made possible for this dream to come true.

Acknowledgments

I would like to first thank God for granting me the gift of life within a wonderful family.

To my parents, José Roberto Polo and Ana Lucia Longo Polo, for their unconditional love, education, and the values that have always guided me. Your dedication and support have been essential at every stage of my life, and especially throughout this academic journey.

To my grandparents, Wilson Longo, Mauriza Gonçalves Longo, Roque Polo, and Maria Fumes Polo, for all the affection, wisdom, and for believing in me even in moments of failure. You are a constant source of inspiration and strength.

To my uncles, Benedito Andrade and Vera Lucia Longo de Andrade, for their companionship and for the support that so often made a difference.

To my sister, Bianca Longo Polo, for her friendship, patience, and encouragement — always ready to listen and support me in everything.

To my friends, especially Fabrício Pelicioni, Pedro Henrique Obara, and Bruno Ramos, who made the journey to ITA lighter, more enjoyable, and unforgettable. Thank you for the laughter, the support during stressful times, and for sharing so many experiences throughout these years.

To my girlfriend, Maria Fernanda Girotto Bertelli, who accompanied me closely through most of my university years, always with love and kindness.

Finally, I thank my professors, who were much more than transmitters of knowledge. Each one, in their own way, contributed to my academic and personal growth, awakening in me critical thinking, curiosity, and a genuine passion for learning.

To all of you, my sincere gratitude for being part of this achievement.

"Success is not final,failure is not fatal:
it is the courage to continue that counts.."
— Winston Churchill

Resumo

Prever com precisão o desempenho real de um voo durante sua fase em rota, especialmente as variações em relação à rota planejada, é fundamental para otimizar o planejamento de combustível e a utilização do espaço aéreo. Este estudo aborda a modelagem preditiva do desempenho operacional em rota no espaço aéreo brasileiro utilizando técnicas de aprendizado de máquina. Desenvolveu-se um modelo de regressão multi-quantílica para estimar o desvio entre a distância realmente voada e a distância planejada durante a fase de voo em rota. O modelo, treinado com o algoritmo CatBoost baseado em árvores de decisão impulsionadas por gradiente, fornece previsões probabilísticas e quantifica a incerteza preditiva. A interpretabilidade local é obtida por meio das Shapley Additive Explanations (SHAP), que oferecem uma compreensão aprofundada da influência relativa das variáveis explicativas. Utilizando um ano de dados operacionais, compostos por informações de vigilância de aeronaves e planos de voo, o método proposto supera abordagens estatísticas de referência, reduzindo o erro multi-quantílico em 77%. Ao integrar técnicas de aprendizado de máquina que combinam precisão preditiva com interpretabilidade, a abordagem proposta busca oferecer um suporte decisório valioso para companhias aéreas e para a gestão do tráfego aéreo, especialmente em áreas como o planejamento de combustível e o gerenciamento de fluxo de tráfego.

Abstract

Accurately predicting the actual performance of a flight during its en-route phase, particularly deviations from the planned flight path, is crucial for optimizing fuel planning and airspace utilization. This study addresses the predictive modeling of en-route operational performance within the Brazilian airspace using machine learning techniques. We develop a multi-quantile regression model to estimate the deviation between the actual flown distance and the planned distance during the en-route flight phase. The model, learned with the CatBoost algorithm based on gradient-boosted decision trees, provides probabilistic forecasts and quantifies predictive uncertainty. Local interpretability is achieved through Shapley Additive Explanations (SHAP), providing insights into the relative influence of explanatory features. Using one year of operational data comprising aircraft surveillance and flight plan information, the proposed method outperforms baseline statistical approaches, reducing the multi-quantile error by 77%. By integrating machine learning techniques that combine predictive accuracy with interpretability, the proposed approach aims to deliver valuable decision support for airlines and air traffic management, particularly in areas such as fuel planning and traffic flow management.

List of Figures

List of Tables

List of Abbreviations and Acronyms

ATM	Air Traffic Management
BRT	Brasília Time
CART	Classification and Regression Trees
DECEA	Department of Airspace Control
GBDT	Gradient Boosted Decision Trees
MMQPE	Mean Multi-Quantile Pinball Error
MPE	Mean Pinball Error
MSE	Mean Squared Error
PE	Pinball Error
SBSP	Congonhas airport
SBRJ	Santos Dumont airport
SHAP	SHapley Additive exPlanations
SIGMA	*Sistema Integrado de Gestão de Movimentos Aéreos*
XAI	ExplainableAI

Contents

1 Introduction

Air transport plays a fundamental role in the territorial, social, and economic integration of Brazil, connecting major urban centers and remote regions in a country of continental dimensions (CATAIA; GALLO, 2007). In this context, the predictability and efficiency of air operations become requirements not only for airlines, which seek to optimize costs and resources, but also for airspace control authorities, which face the challenge of maintaining the safety and efficiency of air traffic flow amid growing demand.

Among the phases of flight, the en-route stage is particularly relevant, as it represents the period during which the aircraft covers the majority of the flight distance. Although the flight plan establishes an ideal trajectory, external factors such as airspace congestion, temporary operational restrictions, air traffic control interventions, and adverse weather conditions often cause deviations that result in significant differences between the planned and the actual trajectory (ZHU *et al.*, 2023). Such variations, even if seemingly small in absolute terms, can have substantial impacts on fuel consumption, emissions, and operational costs.

These uncertainties affect not only airlines and airspace managers but also passengers, who rely on realistic flight time estimates to plan connections and commitments. From an environmental perspective, each deviation that results in an increased distance traveled represents additional fuel consumption and, consequently, higher greenhouse gas emissions, intensifying the challenge for aviation to pursue more sustainable development (TXAPARTEGI; CAZCARRO, 2025).

Figure 1.1 illustrates divergences between the planned trajectory and the actual flown trajectory during the en-route phase for one day of flight operations between Congonhas airport (SBSP) in Sao Paulo and Santos Dumont airport (SBRJ) in Rio de Janeiro, which is one of the busiest routes in the country. Planned trajectories (flight plans) are shown in black and actual trajectories are colored based on the magnitude of the average lateral deviation - DL (in nautical miles) from the planned trajectory (with green lines representing more conforming trajectories and red lines representing less conforming paths). The high frequency of operations on this connection provides a sufficient volume of data for a visual analysis of how difficult it is for the planned and flown trajectories to

coincide, even though the airports are relatively close to each other.

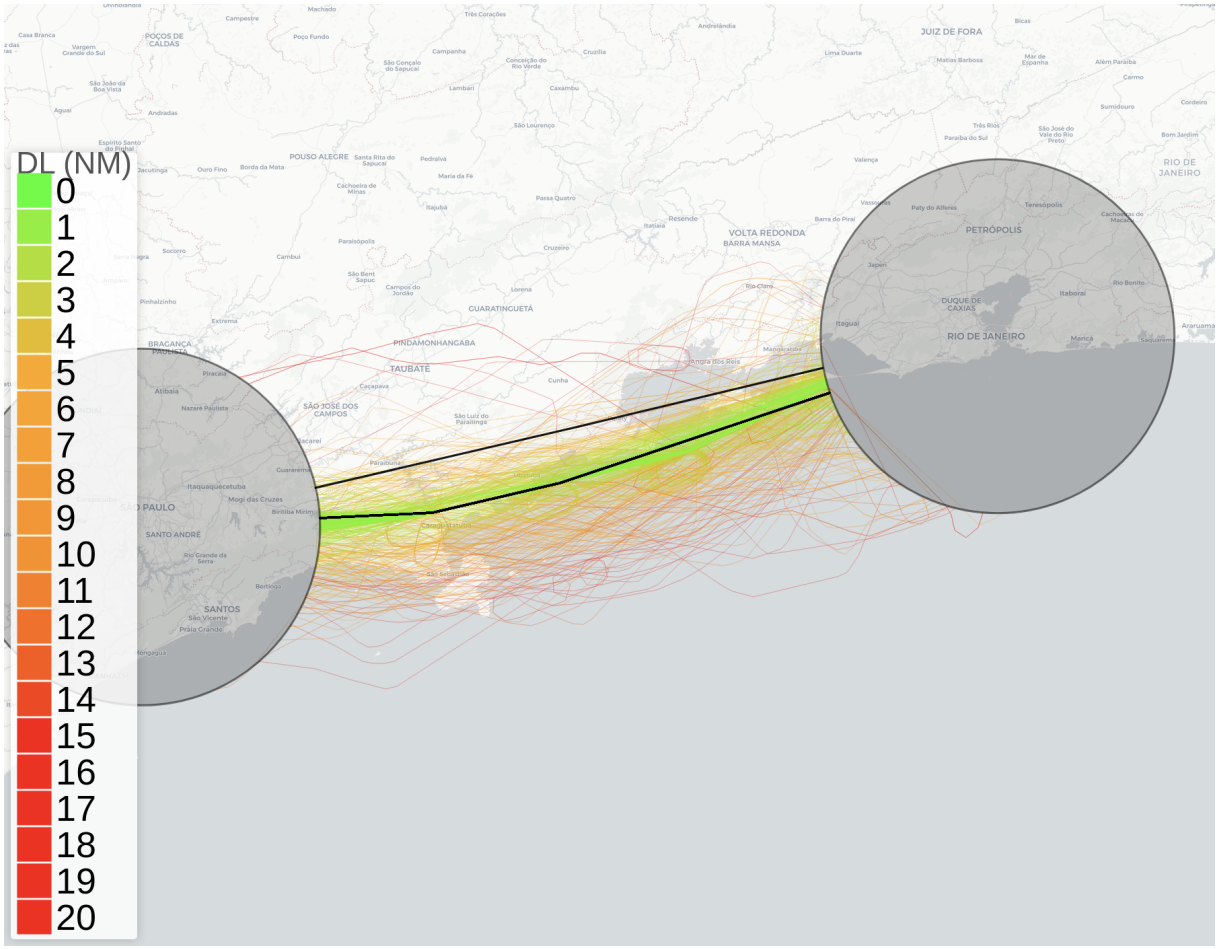


FIGURE 1.1 – Illustration of en-route lateral deviations of actual trajectories from planned trajectories for SBSP-SBRJ.

In this context, accurately predicting en-route operational performance becomes a strategic necessity. By anticipating the magnitude of trajectory deviations, it becomes possible to develop more robust flight plans with improved estimates of time and fuel, while also enabling airlines and control authorities to act proactively.

This study addresses the problem of predictive modeling of en-route performance in the Brazilian airspace. We leverage historical operational data and machine learning methods to develop a prediction model that forecasts the discrepancy between the actual en-route flown distance and the planned distance in the flight plan based on information available at the time of planning. A key feature of our model is that it explicitly quantifies predictive uncertainty through a multi-quantile regression approach. By predicting quantiles instead of a single point estimate, the model provides probable ranges for the deviation between planned and actual flown distance, enabling more robust operational decision-making. Moreover, it is also designed to provide local explanations of predictions towards enhancing its decision support potential for airlines and air traffic management in areas such as fuel

planning and traffic flow management.

As described above, deviations from the originally planned trajectory are often unavoidable due to a range of unpredictable factors, such as air traffic congestion, air traffic control interventions and weather conditions. The application of machine learning techniques can make it possible to anticipate the magnitude of these deviations, as these models are capable of processing large volumes of historical flight data and identifying complex patterns that traditional methods are unable to capture. Furthermore, machine learning stands out for its ability to incorporate a wide range of influencing variables, such as aircraft type, departure and arrival airports, among others. This enables more precise predictions of the actual flight trajectory and total distance traveled, which in turn leads to more realistic and actionable planning.

As a result of this improved planning, one significant advantage is the potential optimization of air operations. With more reliable distance forecasts, airlines can better estimate fuel requirements, optimize flight planning, and ultimately reduce operational costs. Similar predictive strategies have already been applied in other aviation contexts, such as flight delay prediction (DALMAU *et al.*, 2023), where machine learning models analyze multiple variables to anticipate potential disruptions in scheduled operations. By anticipating en-route trajectory deviations from flight plans, traffic managers can more accurately assess downstream airspace utilization - particularly within arrival terminal areas - to improve traffic flow management.

In summary, this study explores the application of machine learning techniques for operational performance prediction towards enhancing the efficiency of flight planning and air traffic management, contributing to improved data-driven decision-making in the aviation sector.

2 Literature Review

The increasing complexity of operational performance in aviation has motivated the use of machine learning techniques to capture variabilities that traditional models fail to address. Dewez *et al.* (2020) propose a statistical framework aimed at estimating aerodynamic variables that are not directly observable, such as drag and lift coefficients, from recorded flight data. The study employs approximate physical relationships to reconstruct these variables and subsequently applies machine learning–based regression algorithms, such as gradient tree boosting, to predict their values under different flight conditions. The proposed approach demonstrates that an aircraft's actual performance can be estimated through data-driven models, enabling more accurate predictions for specific flight phases, such as cruise. This data- and physics-informed approach is also consistent with the objective of the present study, which seeks to predict the variation between the planned and the actual distance flown by an aircraft.

Thiagarajan *et al.* (2017) address the predictability of delays in commercial flights through a two-stage predictive architecture based on machine learning. The developed model begins with a binary classification to determine whether a flight will experience a delay, followed by a regression to estimate the delay time in minutes. Using a large historical dataset containing meteorological and operational information from approximately 3.2 million flights in the United States, the authors explored multiple supervised learning algorithms, highlighting the superior performance of the Extra-Trees Regressor and the Gradient Boosting Classifier.

Based on a machine learning–based approach for flight trajectory prediction, Zhu *et al.* (2024) propose a model aimed at forecasting aircraft behavior under adverse meteorological conditions, specifically in convective weather scenarios. The study highlights the use of a spatiotemporal learning framework combined with a boosting-based ensemble strategy, which emphasizes training samples that reflect explicit deviations caused by weather conditions. This emphasis results in a model capable of predicting trajectories in critical situations for flights between Beijing and Shanghai.

While Zhu *et al.* (2024) focused on predicting trajectories under adverse weather conditions using spatiotemporal models, Liu *et al.* (2021) address the issue of air route in-

efficiency from an explanatory perspective through causal analysis. The study defines inefficiency as the percentage difference between the flown distance and the ideal great circle distance, and proposes a methodological framework composed of trajectory clustering, logit models, and counterfactual analyses. Using this approach, the authors quantify the impact of various factors such as unfavorable winds, restricted military zones, air traffic flow programs, and miles-in-trail restrictions, with special-use airspace accounting for up to 21.9% of the observed inefficiency. This causal decomposition provides valuable insights for the development of more accurate predictive models. Compared to the previous work, which employed deep learning with boosting to capture deviations in convective weather, Liu *et al.* (2021) proposal stands out by offering a structural understanding of the factors leading to suboptimal routes.

In addition to international approaches, national studies have also advanced the analysis of operational efficiency in air traffic. Murça *et al.* (2020) present a data-driven methodology to characterize, at multiple scales, the structure of Brazilian airspace and traffic performance based on flight trajectory data collected from surveillance systems. Using unsupervised learning techniques, such as the Density-Based Spatial Clustering of Applications with Noise algorithm, the authors cluster trajectories in the terminal and cruise phases, defining representative nominal routes. From these routes, structural and operational efficiency metrics are computed, allowing for comparative analyses between different origin–destination pairs. The study also develops statistical models to investigate causal factors affecting efficiency, such as adverse weather, traffic volume, and operational restrictions. This approach provides a foundation for predictive studies by identifying patterns and deviations in actual operations.

Finally, among the various studies analyzed, that of Dalmau *et al.* (2023) represents the greatest convergence with the approach proposed in this paper, both in methodological terms and in operational motivation. The study presents a probabilistic model based on gradient-boosted decision trees, developed to predict departure and arrival delays several days in advance, during the pre-tactical phase, when information such as aircraft rotations or air traffic flow management measures is not yet available. The modeling was applied to Geneva Airport, using exclusively variables accessible within this horizon, such as aircraft type, airline, great circle distance, and estimated passenger load. By predicting the quantiles of the delay distribution instead of single-point values, the authors were able to quantify operational uncertainty. Furthermore, the interpretability analysis reinforced the impact of variables such as the number of passengers and the month of the year on different levels of delay. This work serves as a direct foundation for the modeling developed in the present study, which aims to predict the variation between the planned and the actual distance flown by an aircraft across a range of aerodromes in Brazil.

3 Methodology

This work leverages historical operational data and machine learning techniques for predictive modeling of en-route performance in the Brazilian airspace. A detailed description of the data and methods used is presented in the following sections.

3.1 Data Description and Pre-processing

The data used in this work come from two different sources and cover one year of operations, from January 2023 to December 2023. We used historical flight plan data from the SIGMA system (*Sistema Integrado de Gestão de Movimentos Aéreos*) of the Brazilian Department of Airspace Control (DECEA) to obtain detailed information about planned flight trajectories in the Brazilian airspace. Flight tracking data from FlightRadar24 were used to obtain detailed information of actual trajectories. Data pre-processing was performed to clean, filter and transform the raw datasets into structured datasets of planned and actual flight trajectories segmented by flight phase. Each trajectory was segmented into three phases: terminal area departure, en route, and terminal area arrival. We consider the flight phase segmentation currently adopted for Air Traffic Management (ATM) performance analysis, modeling the departure (arrival) terminal areas with cylindrical volumes with a radius of 40 NM (100 NM) extending from the origin (destination) airports. For each flight, we then computed the planned and actual distances flown during the en-route phase to calculate the performance indicator considered in this work. The final merged dataset used for predictive modeling contains these operational parameters for 673,147 flights, in addition to flight characteristics such as departure airport, arrival airport, airline, and aircraft type.

3.2 En-route Performance Indicator

The choice of en-route performance as our operational performance indicator is based on technical, economic, and environmental factors (PERFORMANCE REVIEW COMMISSION, 2018). This report emphasizes that one of the main reasons for prioritizing en-route effi-

ciency as an indicator lies in its significant impact on fuel consumption, which accounted for about one-third of airlines' operational costs in 2018, a proportion that has been increasing with rising fuel prices, according to the report. Since fuel consumption is directly linked to atmospheric emissions, en-route trajectory efficiency also has a significant relationship with aviation's environmental performance. Therefore, both from an economic and environmental standpoint, improving en-route efficiency becomes a strategic priority.

The performance indicator considered in this work is calculated as the difference between the actual distance flown during the en-route phase and the planned distance in the flight plan. This indicator is slightly different from the trajectory efficiency indicator recommended by the International Civil Aviation Organization (ICAO) for the en-route phase (i.e., KPI 05). Instead of comparing the actual trajectory with an ideal shortest-distance path, we consider the flight plan as a reference. This allows for a more precise quantification of how actual operations diverge from planned operations in order to better support airline and air traffic management decisions regarding fuel planning and traffic flow management.

3.3 Predictive Modeling

The predictive modeling of en-route performance is performed with the application of machine learning techniques on historical data. Murphy (2012) defines machine learning as a set of techniques that enable automatic identification of patterns in the data and the use of these patterns to make predictions and support decision-making under uncertainty. Supervised learning is the machine learning paradigm that focuses on the development of predictive models. In particular, it involves training an algorithm on a labeled dataset, where each training example consists of an input (features) and a corresponding output (target or label). The goal is for the model to learn a mapping from inputs to outputs so that it can accurately predict the output for new, unseen data. Common supervised learning tasks include classification, where the output is a discrete category, and regression, where the output is a continuous value. Algorithms commonly used in supervised learning include linear regression, logistic regression, decision trees, support vector machines, and neural networks.

Our supervised learning problem is framed as a multi-quantile regression problem, as the en-route performance indicator is a numerical variable and we also want to predict the output uncertainty for improved decision support. The variables used as features are presented and explained in Table 3.1. A Gradient-Boosted Decision Tree (GBDT) learning algorithm known as CatBoost is chosen for the supervised learning task. GBDTs have demonstrated strong performance in a wide range of tasks involving tabular data—such as

the dataset analyzed in this study, often surpassing other powerful algorithms, including artificial neural networks. In addition to their predictive accuracy, GBDTs offer advantages such as enhanced interpretability and the inherent capability to manage missing values and categorical variables. As an ensemble learning approach, GBDT constructs a series of decision trees sequentially, with each tree trained to minimize the residual errors of its predecessors using gradient-based optimization techniques. A detailed description of the method and the learning process is presented in the following sub-sections.

TABLE 3.1 – Description of the features used for predictive modeling.

Variable	**Type**	**Description / Examples**
plan_dep	Categorical	Departure airport (SBSP, SBRJ, SBGL, ...)
real_arr	Categorical	Arrival airport (SBRJ, SBGL, SBKP, ...)
day_week	Categorical	Day of the week (1=Sun, ..., 7=Sat)
dep_hour_real	Categorical	Actual departure hour (0–23)
equip	Categorical	Aircraft type (E195, A320, B737, ...)
id_icao_abrev	Categorical	Airline company (GLO, TAM, AZU, ...)

3.3.1 Multi-Quantile Regression

According to Hoffman (2023), when applying machine learning algorithms to regression tasks, it is common for the model to produce a single prediction that represents the most likely value of the target given a set of input features. In general, this prediction corresponds to the mean of the conditional distribution of the output.

However, this approach is limited in scenarios where the data exhibit noise or significant uncertainty, as the expected value does not capture the full complexity of the underlying distribution. Even when the model fits the data well, it does not provide information about the variability of the target around the mean. For example, Figure 3.1 illustrates the described situation.

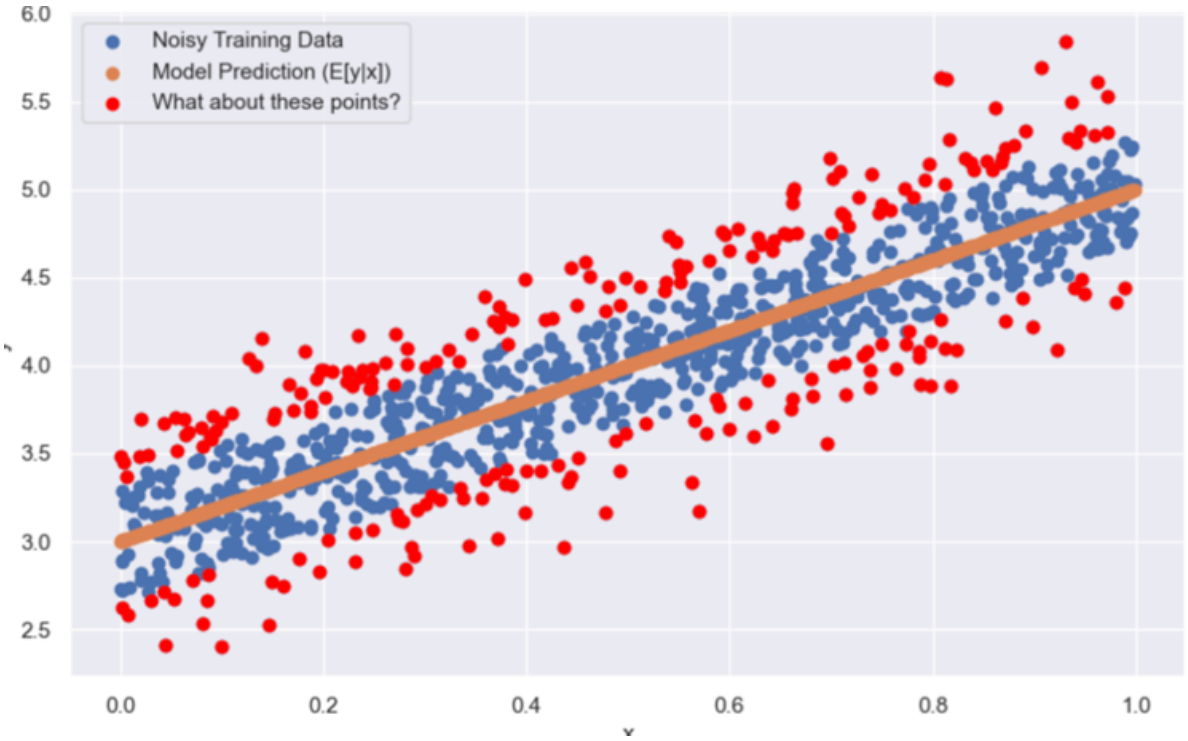


FIGURE 3.1 – Standard regression.

Source: Hoffman (2023).

Quantile regression offers a solution to this problem by modifying the loss function used during model training. Instead of minimizing the Mean Squared Error (MSE), quantile regression uses the Pinball Error (PE), which is asymmetric and depends on the specific quantile being estimated. The PE for a given quantile is defined in Equation 3.1:

$$PE(y, \hat{y}, \alpha) = \alpha.max(y - \hat{y}, 0) + (1 - \alpha).max(\hat{y} - y, 0), \tag{3.1}$$

where α is the quantile, y is the actual output and $\hat{y}$ is the predicted output.

This function penalizes underestimations and overestimations differently. For example, when learning the 95th quantile, the model is penalized by 0.95 for each unit it underestimates the target, and only by 0.05 for each unit it overestimates. This encourages the model to slightly overestimate in order to correctly capture the 95th percentile. The opposite effect occurs when learning quantiles below the median, such as the 5th quantile, where overestimations are penalized more heavily. The loss function that replaces the MSE is the Mean Pinball Error (MPE), given by Equation 3.2.

$$\text{MPE} = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \text{PE}(y_i, \hat{y}_i, \alpha), \tag{3.2}$$

where n_{train} is the the number of training observations.

The conventional quantile regression approach requires training a separate model for each quantile of interest. Since the models corresponding to different quantiles are trained independently, the consistency of the predictions cannot be guaranteed (DALMAU *et al.*, 2023).

By contrast, in multi-quantile regression, a single model is capable of simultaneously predicting multiple quantiles for each observation. This is achieved by optimizing a composite loss function that aggregates the pinball losses of each desired quantile, in order to minimize the Mean MultiQuantile Pinball Error (MMQPE), as given by Equation 3.3:

$$\text{MMQPE} = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \sum_{j=1}^{\alpha} \text{PE}(y_i, \hat{y}_i, \alpha_j). \tag{3.3}$$

This results in an improved model, with a broader view of the data distribution, as illustrated in Figure 3.2.

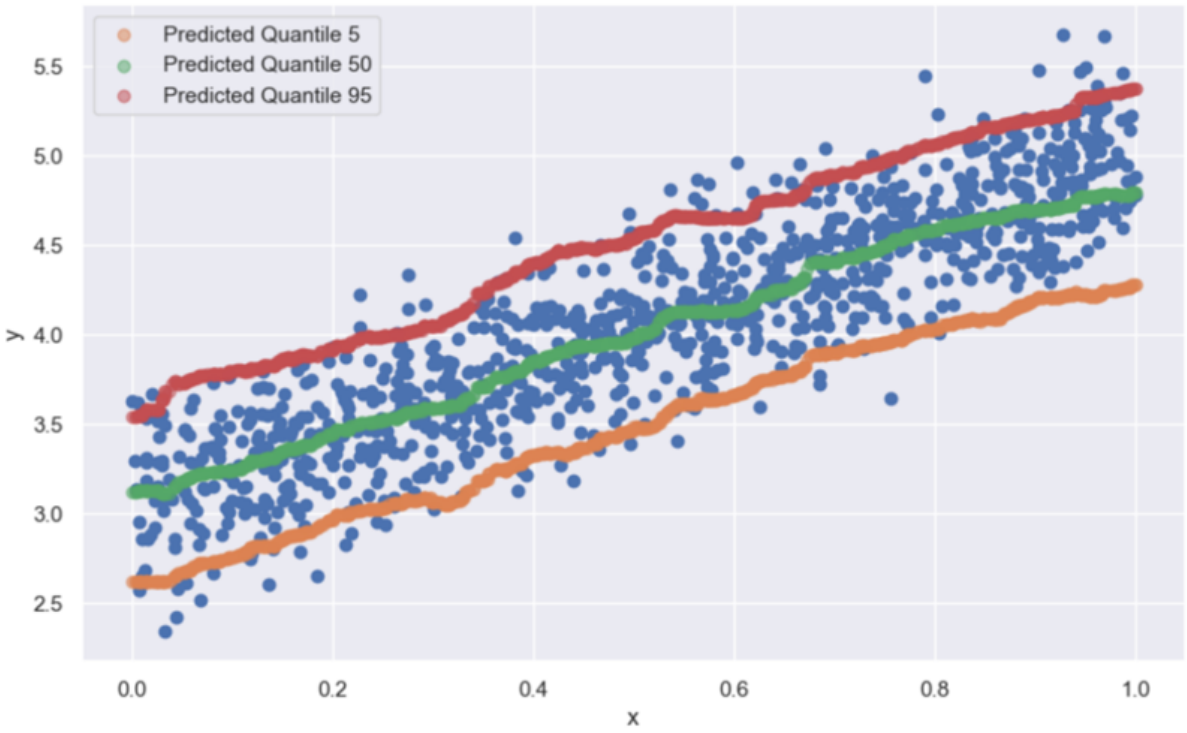


FIGURE 3.2 – Multi-quantile regression.

Source: Hoffman (2023).

Several machine learning algorithms can be used for multi-quantile regression tasks. In this work, we use gradient-boosted decision trees to learn the multi-quantile regression model of en-route performance.

3.3.2 Tree-Based Methods

A tree-based method for classification or regression is based on the sequential partitioning of the space of explanatory variables (FRIEDMAN *et al.*, 2009). One possible representation of this structure is a binary tree, as illustrated in Figure 3.3.

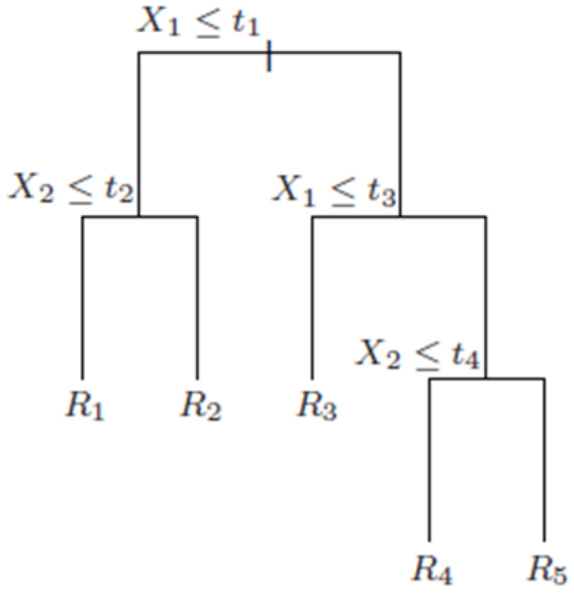


FIGURE 3.3 – Tree example.

Source: Friedman *et al.* (2009).

In this tree, the entire dataset begins at the root node, located at the top. At each internal node, the algorithm evaluates possible splits and selects the one that minimizes the prediction error, typically the sum of squared residuals in the case of regression. Observations that satisfy the splitting condition proceed to the left branch, while the others go to the right branch. This splitting process continues recursively until a stopping criterion is met. The leaves of the tree represent the final regions R_1, R_2, R_3, R_4, and R_5, for example, where each region predicts the response as a constant value optimized locally to minimize the error within that region.

3.3.3 Ensemble Learning

One approach to improve predictive performance is the use of ensemble methods. Rather than relying on a single model to perform a classification or regression task, ensemble methods propose the combination of multiple models that work together to solve the same problem. This strategy is based on the idea that by bringing together several perspectives, even if imperfect, it is possible to reach a more robust and accurate solution than what could be achieved by any individual model alone (ZHOU, 2012).

The distinguishing feature of this approach lies in the interplay between weak and strong models. A weak model is one that, by itself, performs only slightly better than random guessing. In contrast, a strong model has higher accuracy and is capable of generalization, but it is more difficult to build. What ensemble methods demonstrate is that, even when using weak models, it is possible to obtain a strong model by combining them. This gave rise to the concept of ensemble learning.

The construction of an ensemble generally occurs in two stages: generating the base models and combining their predictions. In the first stage, learners are created from the training data with some form of variation to ensure that they learn in different ways. This may include, for example, the use of different subsets of the data, variation in learning parameters, or even the use of distinct algorithms. When all base models belong to the same type, as is the case in this work, where a set of decision trees is used as described in Section 3.3.2, the ensemble is said to be homogeneous. When different algorithms are used, the ensemble is considered heterogeneous. After generating the models, they are combined to produce a single prediction. This combination can be performed using simple mechanisms such as averaging or majority voting, among others. Figure 3.4 illustrates the functioning of an ensemble.

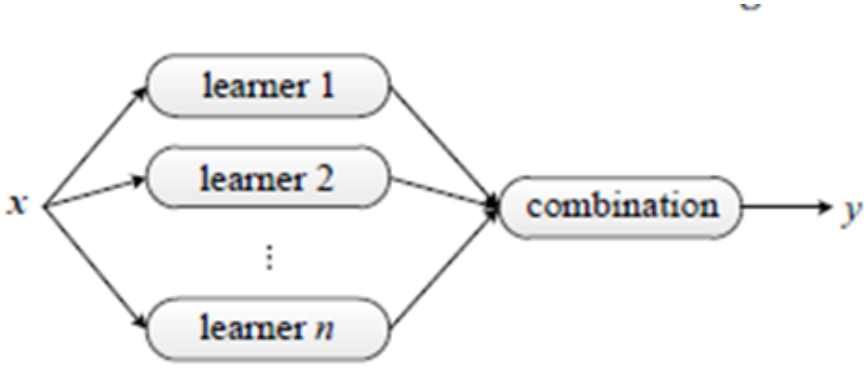


FIGURE 3.4 – A common ensemble architecture.

Source: Zhou (2012).

For an ensemble to perform well, it is not enough for the individual models to be good in isolation; they must also be diverse. Diversity among models is necessary because if all of them make the same errors, there is no benefit in combining them. When the errors occur independently, the chance that the correct predictions of some models compensate for the mistakes of others increases, leading to better overall performance. This diversity can be achieved in several ways, including manipulating the input data, introducing randomness into the training algorithms, or selecting different model architectures.

3.3.4 Gradient Boosted Decision Trees (GBDT)

GBDT method is an ensemble technique that stands out by building additive decision tree models in a sequential manner, where each new model attempts to correct the errors made by the previous one. This approach is grounded in optimization techniques based on gradient descent, applied to the minimization of a loss function over the training data Friedman *et al.* (2009).

The general structure of GBDT is represented as an additive model, as shown in Equation 3.4:

$$f(x) = \sum_{m=1}^{M} \beta_m \, b(x; \gamma_m) \tag{3.4}$$

where $b(x; \gamma_m)$ is a decision tree parameterized by γ_m, and the coefficients β_m determine the contribution of each tree to the final model. The construction follows a process known as forward stagewise modeling, in which components are added one at a time, without modifying those that came before.

At each iteration m, the model is updated using Equation 3.5:

$$f_m(x) = f_{m-1}(x) + \rho_m \, h_m(x) \tag{3.5}$$

where $h_m(x)$ is a new tree fitted to approximate the negative gradient of the loss function, and ρ_m is a learning rate that controls the step size of the update.

Multiple algorithmic implementations of GBDTs exist, with XGBoost, LightGBM, and CatBoost being the most prominent. In this study, CatBoost is chosen as the preferred algorithm because the dataset is largely composed of categorical features, including several with high cardinality. CatBoost offers distinct advantages in such scenarios, as it can natively process categorical variables with minimal preprocessing requirements and effectively reduce the likelihood of overfitting when handling complex categorical data.

3.3.5 Catboost

The learning process in CatBoost begins with a dataset composed of observation pairs, each consisting of feature vectors and their respective outputs, which assume continuous values since this work focuses on a regression task (PROKHORENKOVA *et al.*, 2019).

CatBoost is designed to handle heterogeneous datasets, that is, those containing both categorical variables whose values are names or non-orderable labels (such as ”abbreviations”, ”day of the week”, or ”aircraft type”), and numerical variables, which take on

quantitative values. This distinction is essential because, unlike continuous attributes, categorical ones lack a natural ordering, requiring specific treatment during model training.

Traditional approaches to categorical variables can be inefficient when the variable exhibits high cardinality. For example, a variable like "flight ID", with thousands of unique values, would result in thousands of additional columns after encoding. This significantly increases both computational costs and model complexity.

As a more effective alternative, CatBoost transforms these categories into numerical values based on statistics derived from the target variable, known as target statistics. These aim to estimate the expected value of the target for each observed category.

The main idea behind target statistics is to replace the value of a categorical variable with a numerical estimate of the expected value of the dependent variable, conditioned on the given category. For example, for the category "A321" of the variable "aircraft," the model could replace this category with the average of the target values y observed for all training instances labeled as "A321."

However, the computation of this average, referred to as a "greedy" statistic, introduces a significant statistical issue: a false correlation between the input and the target, known as *target leakage*. This happens because the traditional target statistic includes the target value of the very observation being encoded, which leads to a distortion in the distribution of the training data compared to the test data.

To address this problem, CatBoost introduces a mechanism inspired by an online learning algorithm (PROKHORENKOVA *et al.*, 2019). Instead of using the entire dataset to estimate the target statistic for a given observation, the library simulates an artificial temporal flow by applying a random permutation to the training examples. For each instance x_k, the encoded value of a categorical feature is computed using only the examples that appear before x_k in the permutation. This ensures that the target value y_k of the instance does not influence its own encoding.

For a categorical variable x_i, its encoded value $\hat{x}_{ik}$ in the k-th observation is computed according to Equation 3.6:

$$\hat{x}_{ik} = \frac{\sum_{j:\sigma(j)<\sigma(k)} \mathbb{I}\{x_{ij} = x_{ik}\} y_j + a \cdot p}{\sum_{j:\sigma(j)<\sigma(k)} \mathbb{I}\{x_{ij} = x_{ik}\} + a} \tag{3.6}$$

where σ denotes a random permutation of the observation indices, a is a smoothing hyperparameter, and p is the global mean of the target values y in the training set.

This process prevents information leakage while allowing all data to be used both for training and for computing target statistics.

Additionally, to reduce the variance that may arise in the first elements of the permutation, CatBoost employs multiple independent permutations and computes the encoded values based on them. This stabilizes the estimates and improves the robustness of the final model.

Moreover, as the name suggests, CatBoost is a boosting algorithm. However, it differs from traditional boosting methods where model updates occur through successive gradient adjustments. This difference stems from a problem known as *prediction shift*, which occurs when the gradients are computed using the same model that was trained on the data point itself. This shift arises because the prediction $F^{t-1}(x_k)$, used to compute the gradient at sample x_k, is statistically biased, given that x_k contributed to the construction of the previous model.

This bias causes a cascading effect, as the gradients no longer accurately reflect the direction of the error, the base predictors adjust to incorrect signals, and as a result, the final model suffers from reduced generalization.

To eliminate this issue, CatBoost introduces a mechanism called *ordered boosting*, based on the same ordering principle previously used for handling categorical variables. The algorithm simulates a fictitious temporal flow through a random permutation of the training examples. This ensures that, when computing the gradient for a given sample, only a version of the model trained without that sample is used.

At each iteration, CatBoost aims to fit a new model that reduces the error made by previous predictions. This incremental model is a decision tree constructed to approximate the negative gradient of the loss function, i.e., the direction in which the model needs to be corrected. To do this, the algorithm relies on the gradients computed.

Another distinctive feature of CatBoost lies in the structure of its trees. Instead of employing traditional decision trees with heterogeneous splits at each level, as in standard tree-based methods (see Section 3.3.2), CatBoost uses *oblivious decision trees*, also known as symmetric or balanced trees. In this structure, all nodes at the same level use the same splitting rule, resulting in symmetric and fully balanced trees.

This tree structure offers advantages such as reducing the risk of overfitting and making model application faster and more efficient. Furthermore, the use of uniform splits at each level contributes to model stability, as the impact of small variations in the data tends to be more controlled.

During the construction of each tree, CatBoost evaluates multiple candidate splits and selects those that most effectively approximate the negative gradient. To do this, the quality of each candidate split is assessed based on the cosine similarity between the true gradients and the average gradient values in each leaf. This optimization criterion ensures that the adjustments made are aligned with the errors identified in earlier stages.

In addition, the same tree structure is applied to multiple auxiliary models used in gradient estimation, although the leaf values may differ. That is, all models share the same structural form (i.e., the same splitting rules), but adapt their prediction values according to the portion of the data used in their construction.

More specifically, for each x_i, the value associated with the leaf in which it falls is estimated based on the average of the gradients of the preceding examples (according to the chosen permutation). This value is then used to update the model additively, as expressed in Equation 3.5. This process is repeated for each tree, and in the end, the model prediction is finished.

3.3.6 Supervised Learning Process

For the supervised learning process with the CatBoost algorithm, we split the data into training and test sets, using approximately 80% of the data for training and 20% for testing. During the training phase, hyperparameter optimization was carried out with the aim of enhancing the model's predictive capability without compromising its generalization. For this purpose, a progressive grid search was applied using *HalvingGridSearchCV* (DALMAU *et al.*, 2023), a tool that allows exploring different parameter combinations efficiently, reducing the computational cost during the process.

3.3.6.1 Search Space

In the training phase, three essential CatBoost hyperparameters were selected for tuning: `depth`, `learning_rate`, and `iterations`. The first relates to the complexity of the generated trees, while the second controls the step size in the gradient optimization process. The `iterations` parameter defines the maximum number of boosting stages and is used here as a resource by the halving method. The initial search space was defined as shown in Table 3.2:

TABLE 3.2 – Parameters explored during the optimization process.

Hyperparameter	**Values**
Tree Depth (`depth`)	[4, 6, 8]
Learning Rate (`learning_rate`)	[0.05, 0.1, 0.2]
Iterations (`iterations`)	[200]

The selection of these values aimed to balance fitting capacity and computational cost, avoiding both *underfitting* and *overfitting*. The values found by the model are shown in Table 3.3:

TABLE 3.3 – Parameters obtained during the optimization process.

Hyperparameter	Values
Tree Depth (`depth`)	8
Learning Rate (`learning_rate`)	0.2
Iterations (`iterations`)	200

3.3.6.2 Temporal Cross-Validation

Considering that the data exhibit temporal dependence in flight operations, segmented time cross-validation (*TimeSeriesSplit*) was chosen. Unlike traditional cross-validation, this approach prevents future observations from being used to train the model, ensuring that the process respects the true chronological order of events, avoiding data leakage and over-optimistic performance estimates (JONES, 2025). For this purpose, the training set was divided into 10 subsets along the time series.

3.3.6.3 Progressive Resource-Based Search

For the optimization step, the *HalvingGridSearchCV* method was used, the same as employed by Dalmau *et al.* (2023), which initially runs all models with a smaller amount of resources (reduced number of observations). In each round, only a fraction of the best-performing models is selected, while the amount of data used is progressively increased. This combines a wide initial search with computational efficiency.

As the evaluation metric, the MMQPE was adopted, defined as the MPE calculated for the quantiles $\{0.05, 0.25, 0.50, 0.75, 0.95\}$. In this way, the optimization process considered the overall model performance across the entire conditional distribution of route deviations, rather than just at a specific point of the forecast.

At the end of the iterations, the model with the best MMQPE performance was selected as the final version for application to the prediction set.

3.4 Model Explanation

The growing use of more complex predictive models in recent years has increased the need for tools that enable local interpretation of predictions. Explainability plays a crucial role in fostering trust in machine learning models and ensuring their effective adoption by end users for informed decision-making. In recent years, the concept of Explainable AI (XAI) has gained significant attention, reflecting the growing need to make complex models more transparent, interpretable, and accountable.

Towards this goal, one possible approach is SHapley Additive exPlanations (SHAP), introduced by Lundberg *et al.* (2020). The method is based on game theory principles, as developed by Lloyd S. Shapley, particularly the Shapley values, to assign each input variable a quantitative importance associated with its contribution to the model's prediction.

The essence of SHAP lies in defining an additive explanatory function, where the prediction of a complex model $f(x)$ is approximated by a weighted sum of the individual contributions of each variable. The additive explanatory function is given by Equation 3.7:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.7}$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

where z' is a binary vector representing the presence (1) or absence (0) of a given feature in the explanation, and ϕ_i represents the Shapley value. The coefficients ϕ_i are determined to satisfy three desirable properties: local accuracy, missingness, and consistency.

- ***Local accuracy*** requires that the sum of the contributions attributed to all features equals the model prediction for a specific input x. According to Equation 3.8:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{3.8}$$

  where $f(x)$ is the original model's output for the input instance x. This property ensures that no systematic error is introduced, preserving the model's predicted value.

- ***Missingness*** states that if a variable is not present in the explained input instance (or is marked as absent in the simplified vector x'), then its contribution to the explanation must be exactly zero. According to the Equation 3.9:

$$if \quad x'_i = 0, \quad then \quad \phi_i = 0 \tag{3.9}$$

  This condition ensures that absent features do not influence the explanation, in other words, the method accounts for their presence or absence.

- ***Consistency*** establishes that if the contribution of a given variable to the prediction increases (or remains the same) in all possible scenarios between two models f and f', then the value assigned to that variable cannot decrease.

  If for every subset of variables $z' \in \{0,1\}^M$, the inequalities 3.10 and 3.11 are true:

$$f'(z') - f'(z' \setminus i) \geq f(z') - f(z' \setminus i) \tag{3.10}$$

  then

$$\phi_i(f') \geq \phi_i(f) \tag{3.11}$$

  This property means that if the influence of a feature increases from one model to another, its assigned importance must reflect that increase, ensuring consistency in explanations.

Based on these three properties, Lundberg *et al.* (2020) demonstrate that there is a unique way to assign the ϕ_i values that is compatible with the model definition. This assignment corresponds exactly to the Shapley values formula from game theory, given by Equation 3.12:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus \{i\}) \right] \tag{3.12}$$

where $|z'|$ is the number of non-zero entries in z', and z' $\subseteq$ x' represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'.

However, the exact calculation of the Shapley values shown in Equation 3.12 becomes computationally infeasible for models with many variables, as is the case in this work, as shown in Section 3.1, since it requires evaluating the model across all possible combinations of variables (complexity 2^M). To overcome this limitation, an approximation method was used: TreeExplainer, which reduces the complexity of calculating exact Shapley values from exponential to polynomial time, while still providing explanations with theoretical guarantees of accuracy and consistency. TreeExplainer is optimized for decision tree–based models, as also done by Dalmau *et al.* (2023).

After applying this model, it was possible to carry out an analysis that allowed local interpretation of specific predictions using graphs such as the waterfall plot, which displays each feature's individual contribution to each quantile separately. This visualization is particularly useful in quantile regression models, where interpreting different quantiles can

reveal information that is not captured by the global model, as it represents an average.

3.5 Model Assumptions and Limitations

The predictive model developed in this study is based on a set of methodological assumptions and operational simplifications that must be taken into account when interpreting the results. First, it is assumed that the historical behavior observed in the training dataset is representative of future operational conditions. Therefore, the statistical relationships learned by the model are expected to remain stable over time. Although this assumption is common in predictive studies, such as those referenced in Section 2, it may not hold under structural changes in airspace configuration, airline networks, or air traffic management policies.

Another important assumption concerns the set of variables used. The model relies exclusively on operational and categorical features available before takeoff, such as origin and destination airports, aircraft type, airline, day of the week, and departure time. It is assumed that these variables capture the main sources of variability associated with en-route route deviation. However, relevant external factors are not represented, particularly meteorological conditions (e.g., wind patterns or convective weather) and temporary operational constraints (such as airspace or sector closures). The absence of such elements represents a limitation, as deviations in the flight trajectory often arise directly from these non-deterministic influences.

Regarding the modeling technique, the CatBoost multi-quantile method ensures consistent probabilistic forecasts; however, it does not guarantee physical interpretability of the mechanisms that drive route deviations. Although the SHAP analysis provides feature-level explanations, these correspond to statistical associations rather than causal relationships. The model recognizes patterns based on correlations in the data, without explicitly accounting for aerodynamic, meteorological, or air traffic flow dynamics.

Another relevant aspect is the imbalance within the dataset across operational segments. Some airports, aircraft types, airlines, and departure time brackets have significantly more samples than others. As a result, both overfitting and underfitting may occur in different regions of the dataset.

Finally, the model assumes independence between flights, disregarding network effects such as reactionary delays or dynamic traffic adjustments stemming from widespread congestion. In real-world operations, route deviations may propagate across multiple flights due to coordinated air traffic management initiatives, a phenomenon that is not explicitly captured by the model.

4 Results and Discussion

This section presents an exploratory analysis of the data and discusses the performance of the predictive model developed in this study.

4.1 Exploratory Data Analysis

Before training the predictive model, an exploratory analysis was conducted to better understand the main characteristics of the dataset. This analysis was carried out with the support of data visualizations.

Figure 4.1 shows the main departure airports in the data, which are associated with a higher concentration of flight operations.

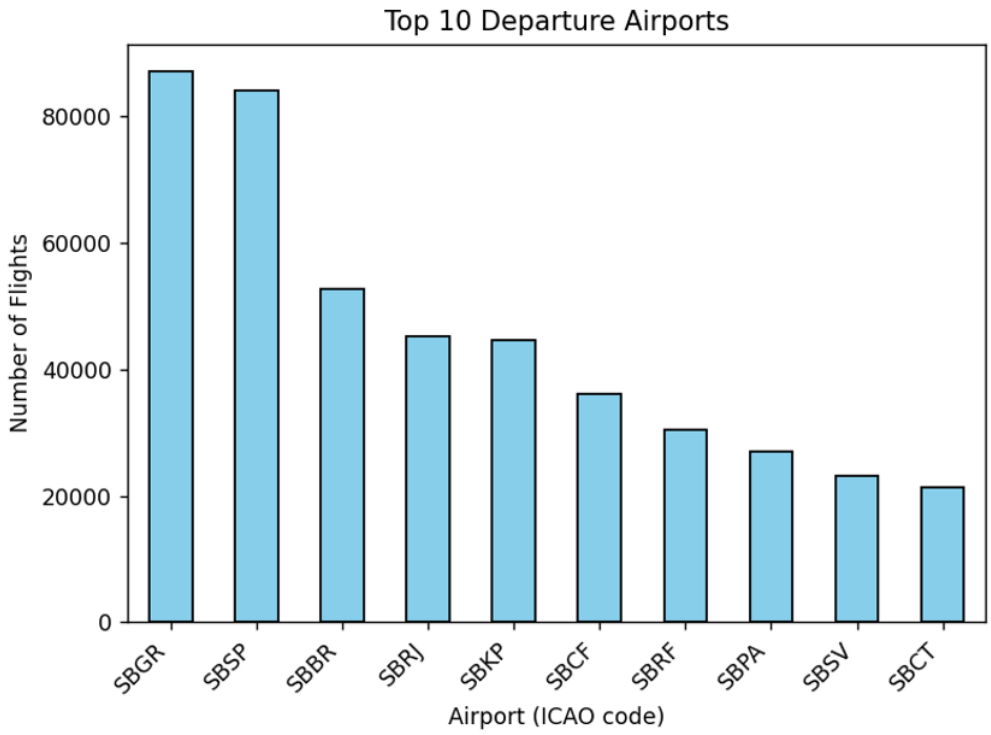


FIGURE 4.1 – Top 10 departure airports.

Figure 4.2 shows the main aircraft models used in the analyzed routes.

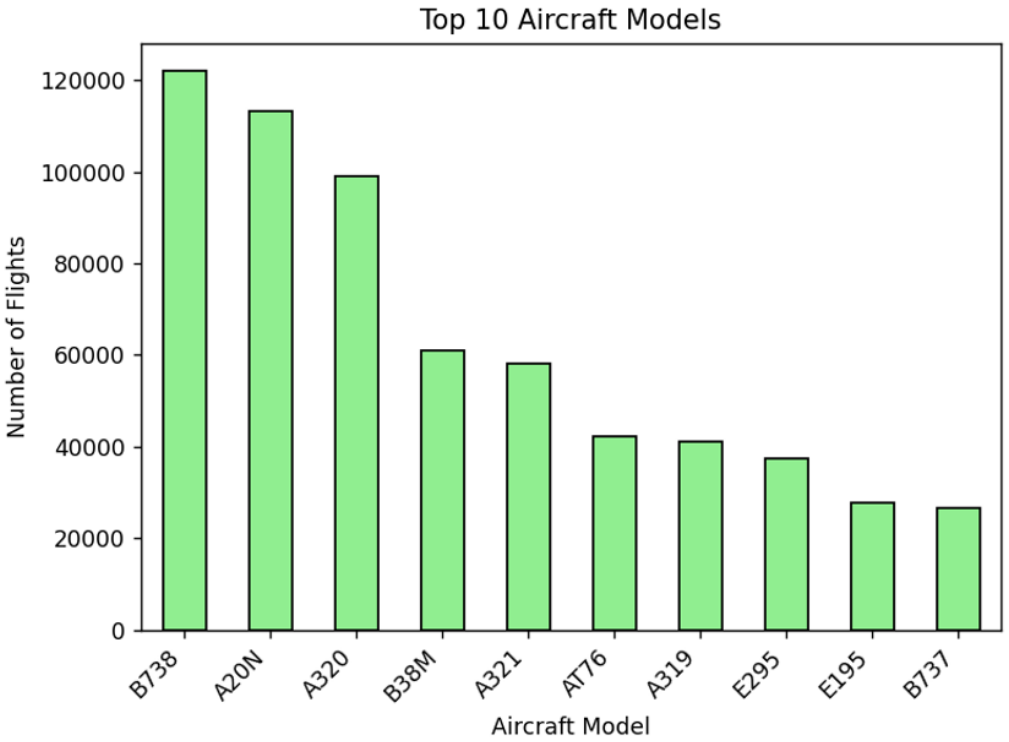


FIGURE 4.2 – Top 10 aircraft models.

Figure 4.3 presents the four major airline operators, with TAM, GOL, and AZUL being the only ones with a significant number of flights, while the others operate a considerably smaller volume.

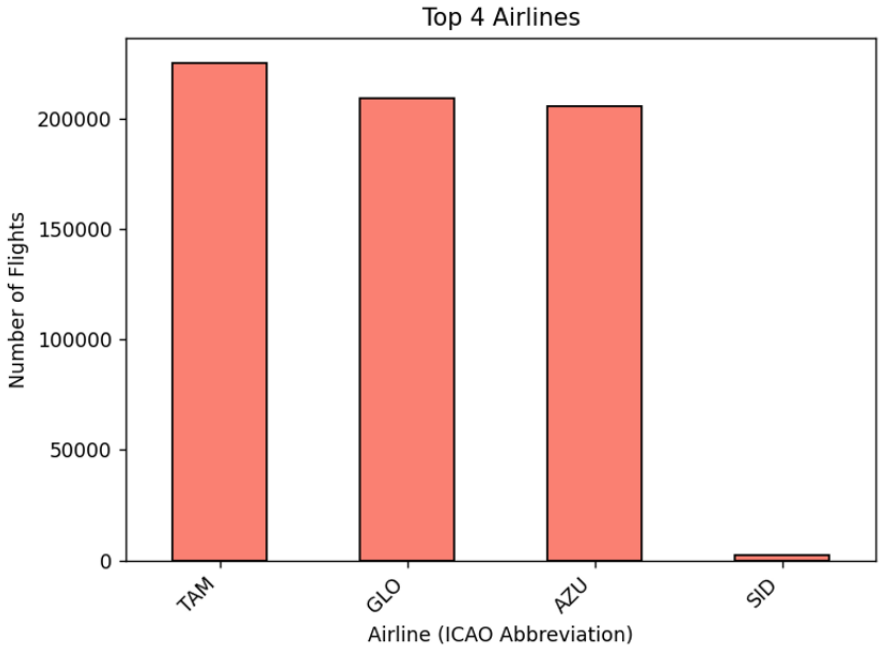


FIGURE 4.3 – Top 4 airlines.

Figure 4.4 displays the time of day with the highest concentration of simultaneous

flights, indicating that the morning period is less busy compared to the others.

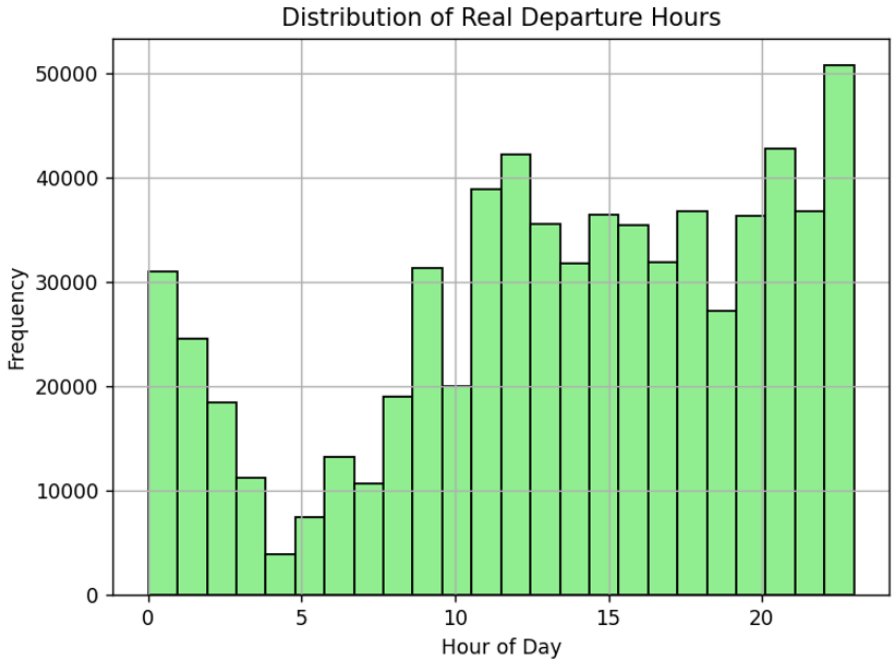


FIGURE 4.4 – Distribution of actual departure hours.

Finally, Figure 4.5 shows the number of flights by day of the week, from which it can be concluded that there is a noticeable reduction in flight activity during weekends.

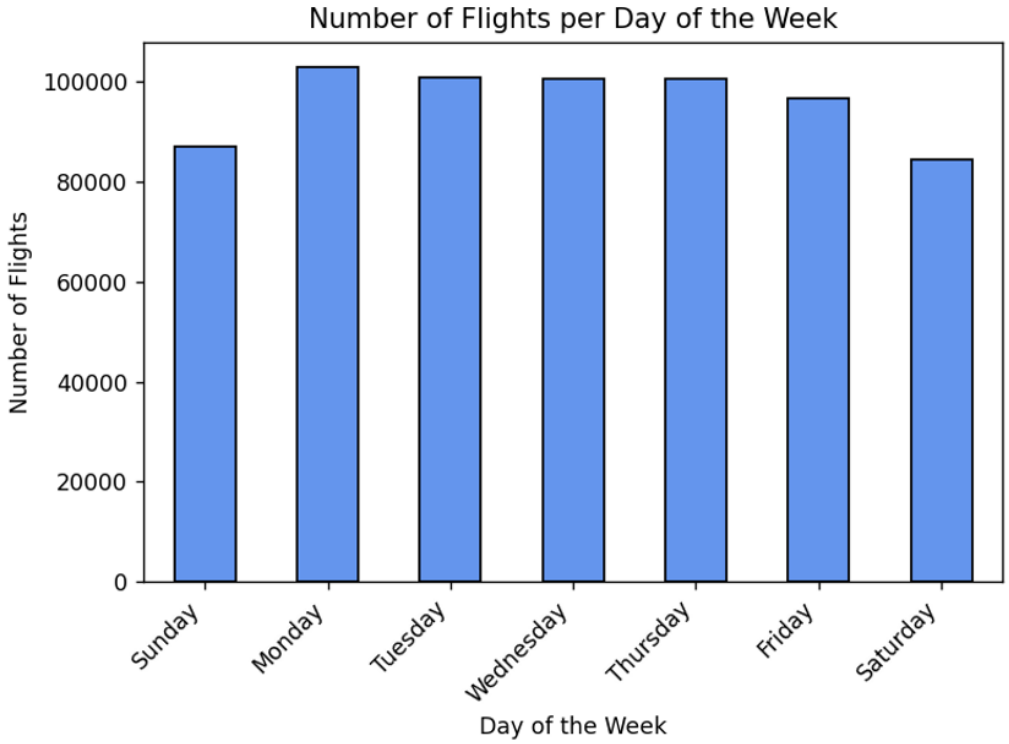


FIGURE 4.5 – Number of flights per day of the week.

In order to gain a better understanding of the target variable, which represents the

difference in trajectory distance between the planned and the actual path flown by the aircraft, flights were observed during certain periods of the day. For this, a division of the day into specific time periods was made, as shown in Table 4.1.

TABLE 4.1 – Classification of time periods throughout the day.

Time of day	**Start (BRT)**	**End (BRT)**
Early Morning	02:00	06:00
Morning	06:00	12:00
Afternoon	12:00	17:00
Evening	17:00	21:00
Night	21:00	23:00
Late Night	23:00	02:00

Based on a reorganization of the dataset according to Table 4.1, it was possible to create cumulative plots of the target variable for each time of day. These plots are presented in Figures 4.6, 4.7, 4.8, 4.9, 4.10, and 4.11, providing an overview of how the quantiles behave throughout the day.

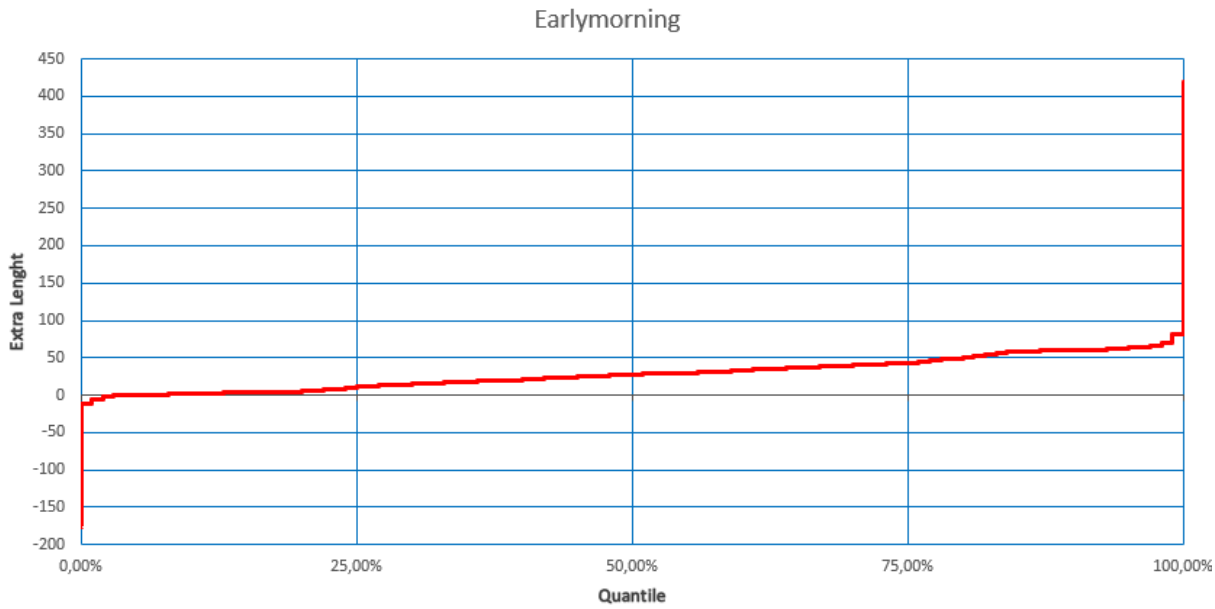


FIGURE 4.6 – Cumulative distribution of horizontal en-route deviation for early morning flights.

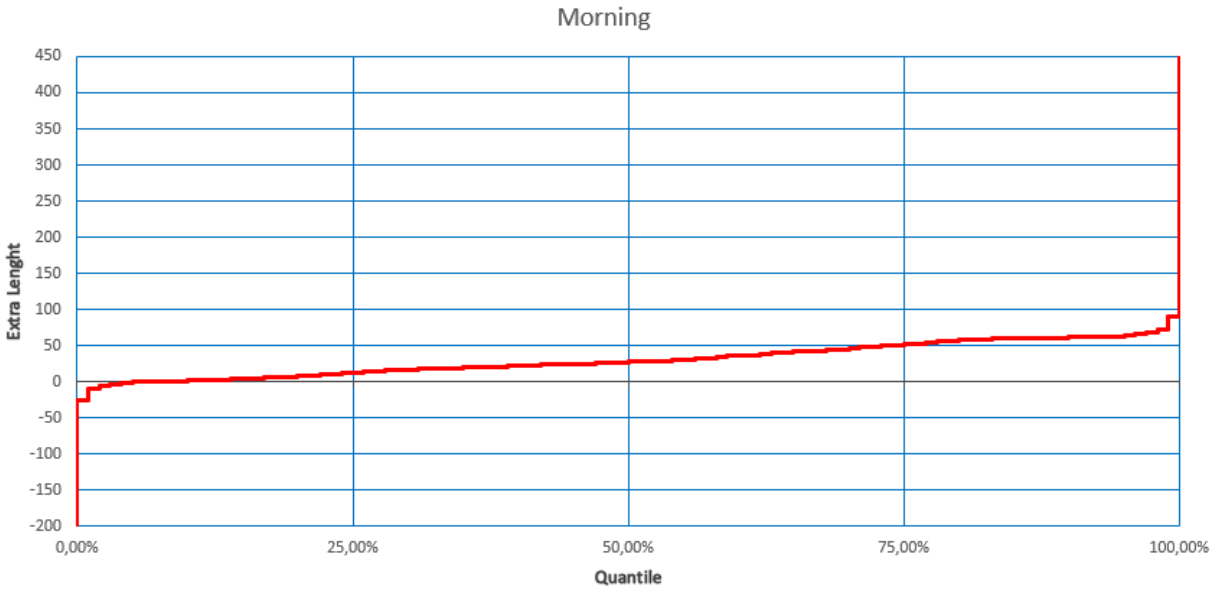


FIGURE 4.7 – Cumulative distribution of horizontal en-route deviation for morning flights.

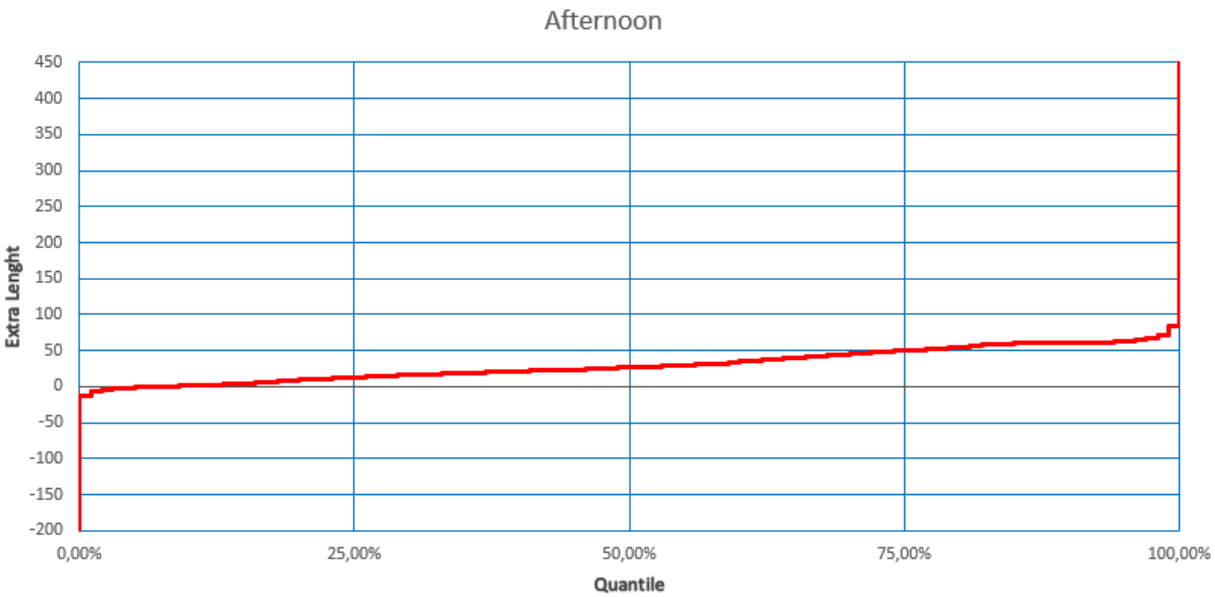


FIGURE 4.8 – Cumulative distribution of horizontal en-route deviation for afternoon flights.

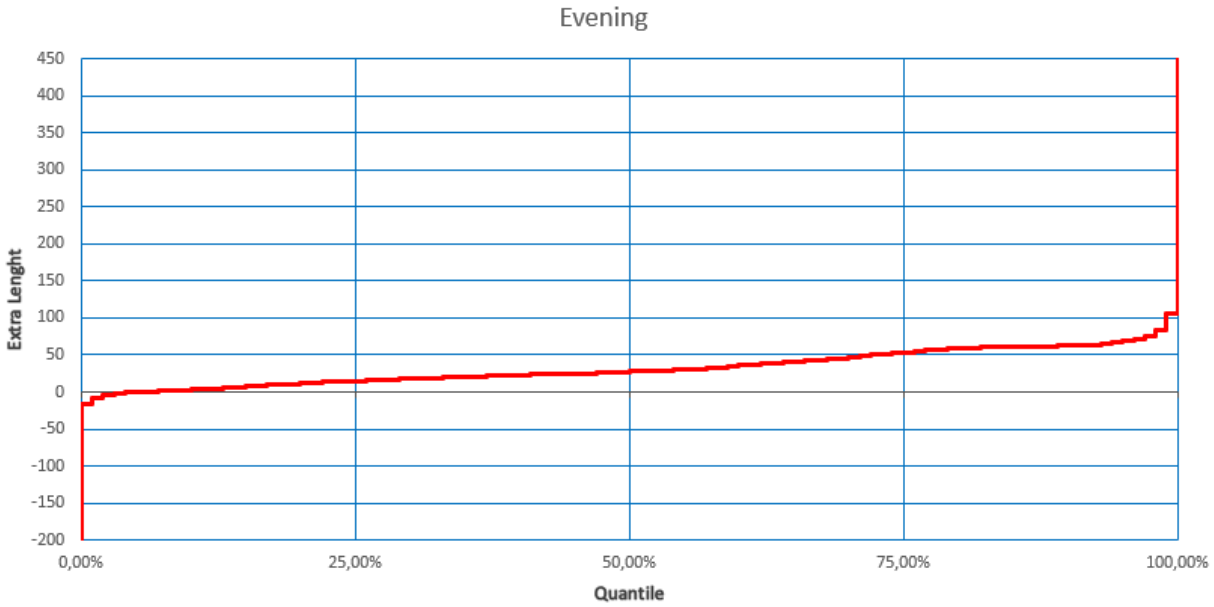


FIGURE 4.9 – Cumulative distribution of horizontal en-route deviation for evening flights.

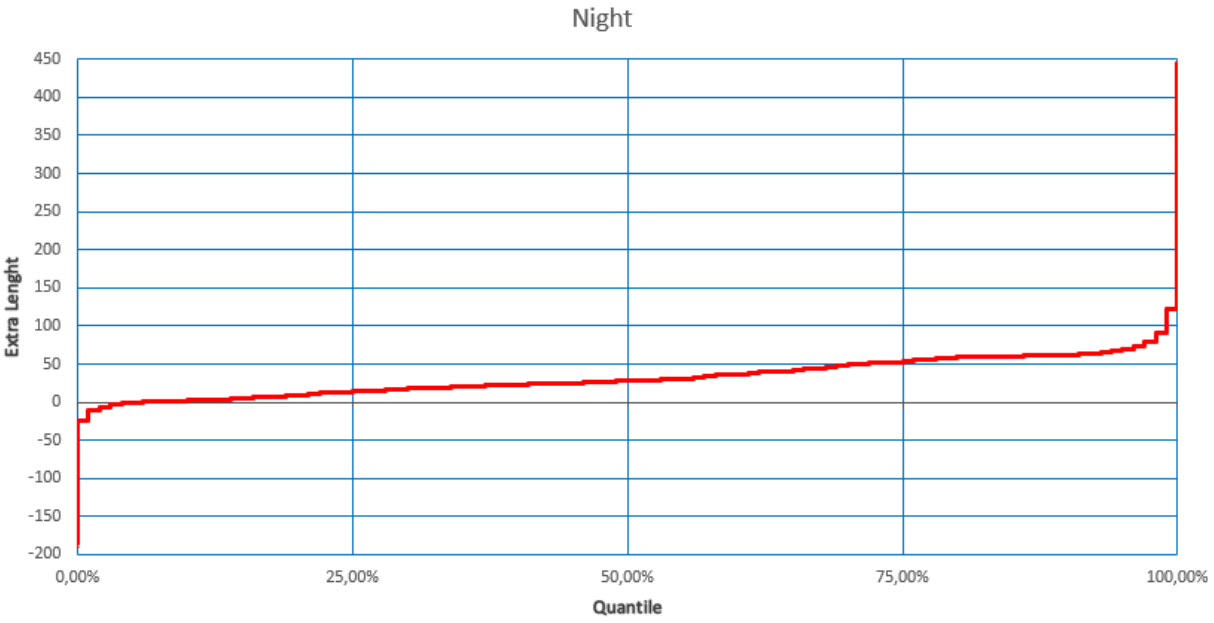


FIGURE 4.10 – Cumulative distribution of horizontal en-route deviation for night flights.

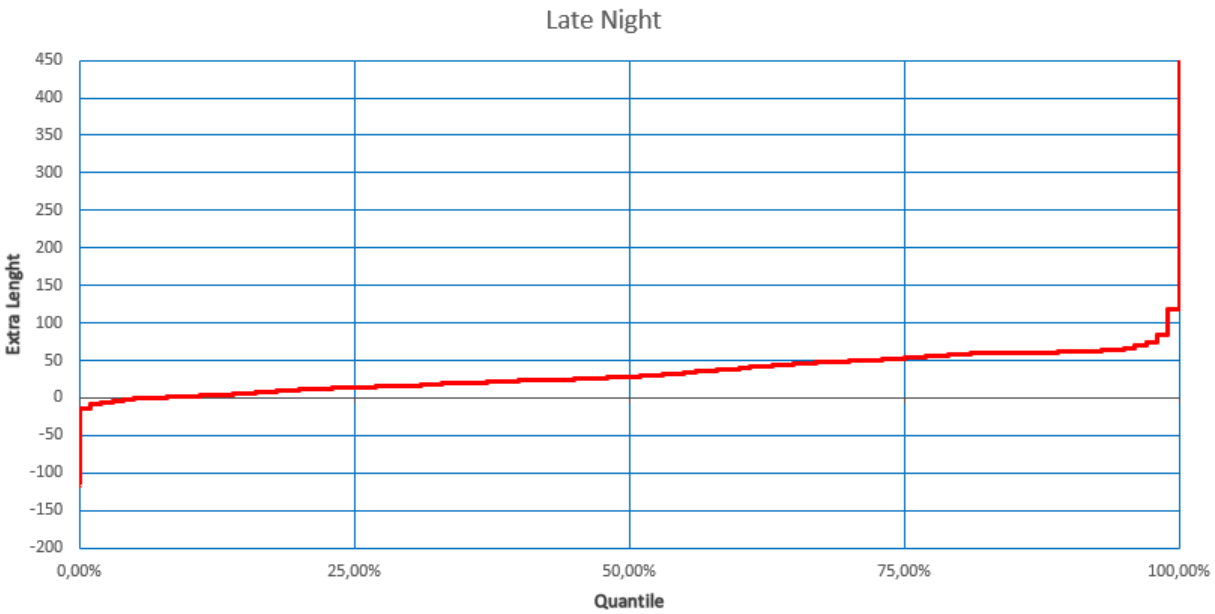


FIGURE 4.11 – Cumulative distribution of horizontal en-route deviation for late night flights.

In order to enable a later comparison between the quantiles predicted by the machine learning model, as described in Section 3.3.1, and the values from the original dataset, Table 4.2 was constructed.

TABLE 4.2 – Quantiles of horizontal en-route deviation by time of day (in nautical miles).

Quantile	Early Morning	Morning	Afternoon	Evening	Night	Late Night
5%	0.14	-0.85	-0.67	-0.03	-0.65	-0.74
25%	10.77	12.92	13.26	14.55	13.92	13.77
50%	27.61	27.43	26.56	27.25	28.26	28.67
75%	43.18	51.77	50.24	53.13	53.30	52.80
95%	62.69	63.33	62.87	67.45	68.77	65.74

4.2 Model Predictive Performance

The multi-quantile regression model learned with CatBoost was applied in the test dataset - composed of 134,630 flights - to predict the extra en-route flight distances given operational conditions. The predictive performance of the model was evaluated using the MPE for each quantile, as well as the MMQPE, which provides a global measure of predictive accuracy according to Section 3.3.1.

To quantify the impacts of the machine learning approach, we compare the predictions generated with the CatBoost model with baseline predictions derived from basic statistical principles, often used in current practice. To build our baseline prediction model, observations in the training dataset were grouped by period of the day and the quantiles

were calculated for the performance distribution of each group, as shown in Table 4.2. Then, for each observation in the test set, the quantile distribution is predicted using the quantiles computed for the group into which it falls.

Table 4.3 presents the quantile error and the overall multiquantile error in the test set for the predictions made with both methods. It is observed that the CatBoost model demonstrated superior performance compared to the baseline model for all evaluated quantiles. While the baseline exhibited significantly higher errors, especially in the extreme quantiles (5% and 95%), CatBoost maintained substantially lower MPE values, indicating greater accuracy in the predictions.

Furthermore, the aggregated MMQPE value reinforces this superiority: the CatBoost model achieved an average error of 4.92, representing a significant reduction compared to the baseline model error of 21.70. This difference shows that the CatBoost model not only improves point accuracy for specific quantiles but also provides greater overall consistency across the full range of predictions.

Therefore, it is concluded that the machine learning approach with CatBoost proved to be more efficient for the problem under study, being able to better capture the variability of the data and provide more reliable estimates for different quantile levels.

TABLE 4.3 – Quantile and multi-quantile errors for the CatBoost and baseline prediction models of en-route performance.

Quantile	**MPE (Baseline)**	**MPE (CatBoost)**
5%	33.48	2.03
25%	17.76	5.19
50%	10.07	6.77
75%	16.70	7.21
95%	30.50	3.40
MMQPE	**21.70**	**4.92**

Besides the calculated error, the 5% (q_5) and 95% (q_{95}) quantiles were considered as boundaries of a wide confidence interval within which the actual trajectory deviation value is expected to lie. Thus, for each prediction instance, the following condition in Equation 4.1 was verified:

$$q_5 \leq \text{actual value} \leq q_{95} \tag{4.1}$$

If this condition was satisfied, the model was considered to have achieved good coverage for that specific observation. The proportion of correct predictions over the total number

of observations was then computed, resulting in the metric known as *Coverage*, defined by Equation 4.2:

$$Coverage = \frac{\text{Number of correct predictions}}{\text{Total observations}} \times 100\% \tag{4.2}$$

This metric enables the assessment not only of point wise forecast accuracy but also the model's ability to appropriately represent the distribution of deviations, capturing both extreme cases and operational uncertainties.

A total of 91,387 correct predictions were found out of 134,630 in the data test, resulting in a global coverage value of 67.8%. This coverage indicates that the model successfully captured approximately two-thirds of the actual observations within the predicted quantile intervals. This result suggests that the model provides a reasonable representation of the data's variability, though it tends to slightly underestimate the underlying uncertainty. Future adjustments to the quantile configuration or model hyperparameters could further improve the calibration and bring the coverage closer to the desired level.

4.3 Explainability through SHAP Analysis

When a machine learning model is used to generate predictions that support decision-making, it is essential to clearly communicate the specific context in which each prediction is made. Providing this contextual understanding helps build trust in the model's outputs and ensures that decisions based on them are well-informed and reliable. We leverage the explainability power of the GBDT models to provide explanations for en-route performance predictions. For this, we use SHAP, as described in Section 3.4.

To illustrate the explanainability layer of our approach, we selected four instances of our test dataset in order to observe how the feature *hour* influences the final prediction. The criteria adopted for case selection were primarily based on choosing the two busiest arrival airports in Brazil, namely SBGR and SBSP. For each airport, representative times of low and high air traffic were defined. As illustrated in Figure 4.4, the 08:00 time slot shows a lower volume of operations, while 23:00 corresponds to a period of higher traffic, and was therefore chosen for comparison. Regarding the day of the week, Monday was selected as it is the day with the highest number of flights, as indicated in Figure 4.5. With respect to the aircraft and airline, the same equipment type and operator were selected across the cases to minimize potential interference in the prediction, since the goal is to observe the influence of the feature *hour*. The selected cases are presented in Table 4.4.

TABLE 4.4 – Selected instances for en-route performance prediction explanation.

Departure	Arrival	Day	Hour	Aircraft	Airline	Deviation (NM)	Q=95% (NM)
SBBR	SBGR	2	8	A321	TAM	65.98	89.25
SBBR	SBGR	2	23	A321	TAM	70.11	107.94
SBBR	SBSP	2	8	A320	TAM	11.80	68.93
SBBR	SBSP	2	23	A320	TAM	11.06	85.03

Figures 4.12, 4.13, 4.14, and 4.15 present the SHAP explanations for each instance in Table 4.4. The SHAP explanations are represented using a waterfall plot for each individual instance for the 95% quantile. In this plot, the feature values for the selected instance are displayed on the left side. Features are arranged based on their corresponding SHAP values, which quantify their contribution to the model's prediction. A blue arrow signifies that a feature decreases the predicted value, whereas a red arrow indicates that a feature pushes the prediction toward higher values.

It is observed that during periods of lower air traffic volume, such as at 08:00, the SHAP value for the feature *hour* is negative, indicating that it contributes to reducing the trajectory deviation relative to the planned flight path. Conversely, at 23:00, which corresponds to a period of higher traffic volume, the SHAP value becomes positive, suggesting an increase in the predicted trajectory deviation. For instance, in the case of SBSP airport, the feature *hour* reduced the deviation from the base value by approximately 4.72 NM at 8:00, but increased it by about 4.42 NM at 23:00. Therefore, it is evident that during periods of higher demand volume, the model tends to predict larger trajectory deviations. This is also evident in Table 4.5, which shows the average SHAP values for the feature *hour* at different times of the day across all arrival flights at SBGR and SBSP.

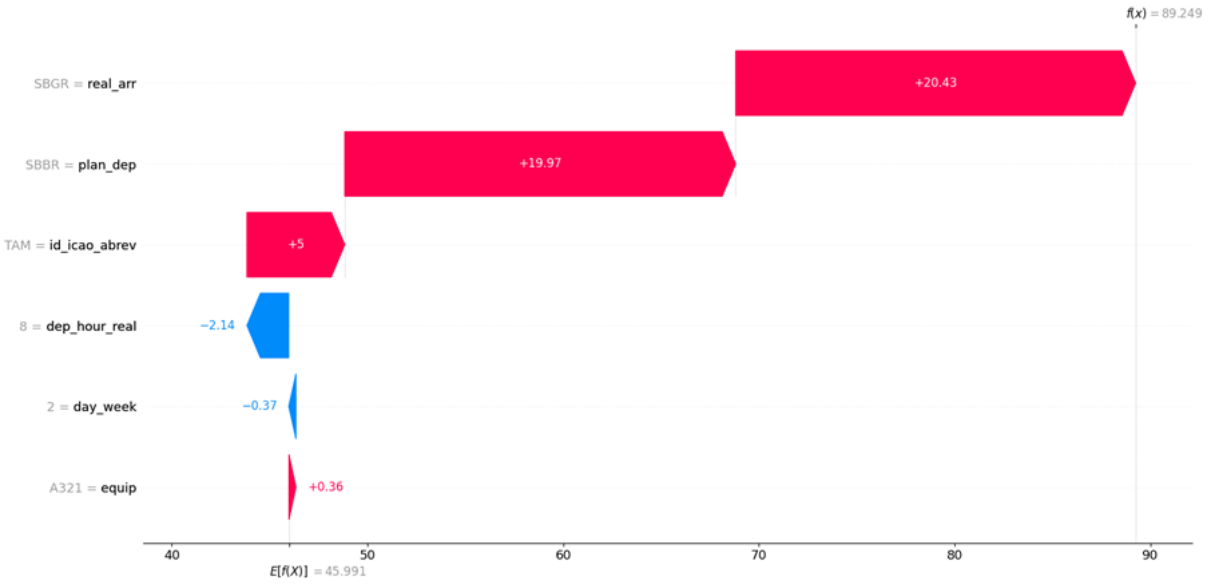


FIGURE 4.12 – SHAP explanations for an arrival flight at SBGR at 8:00.

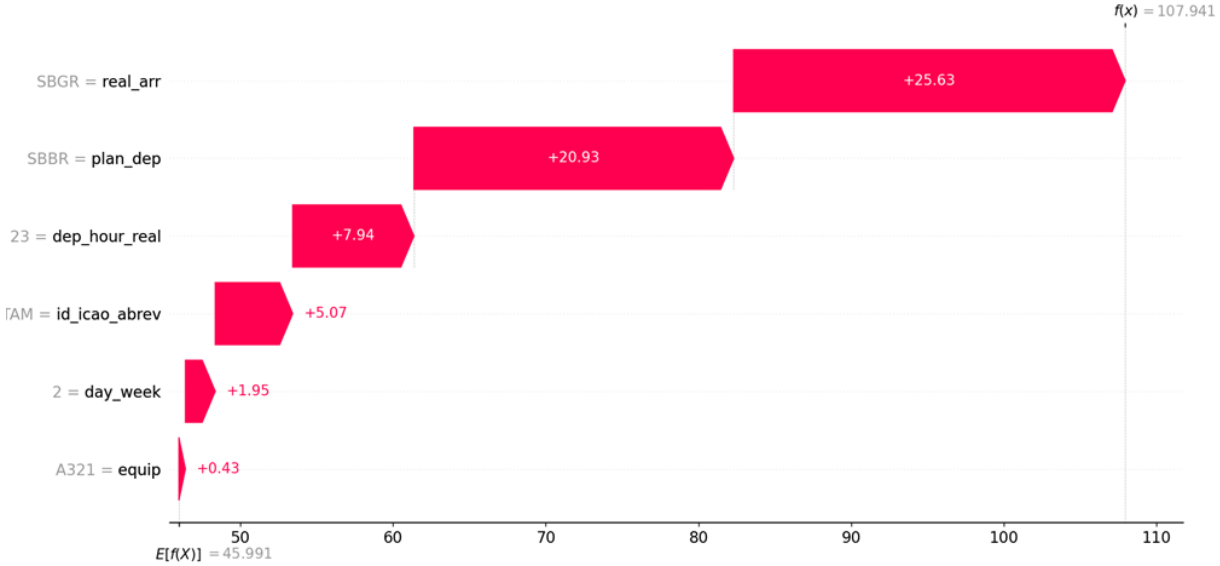


FIGURE 4.13 – SHAP explanations for an arrival flight at SBGR at 23:00.

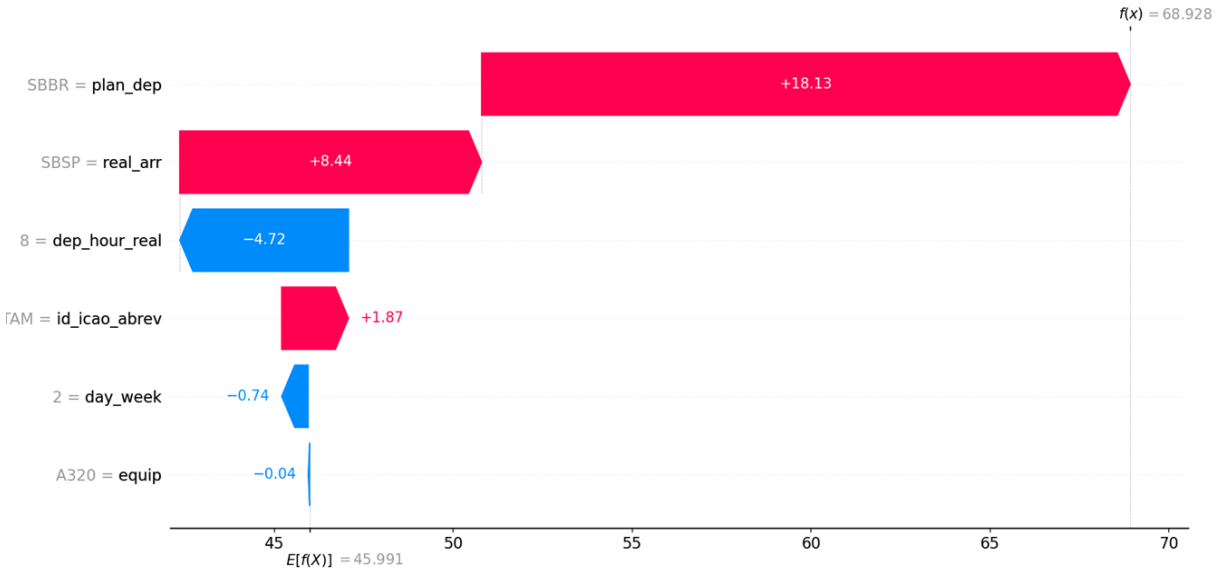


FIGURE 4.14 – SHAP explanations for an arrival flight at SBSP at 8:00.

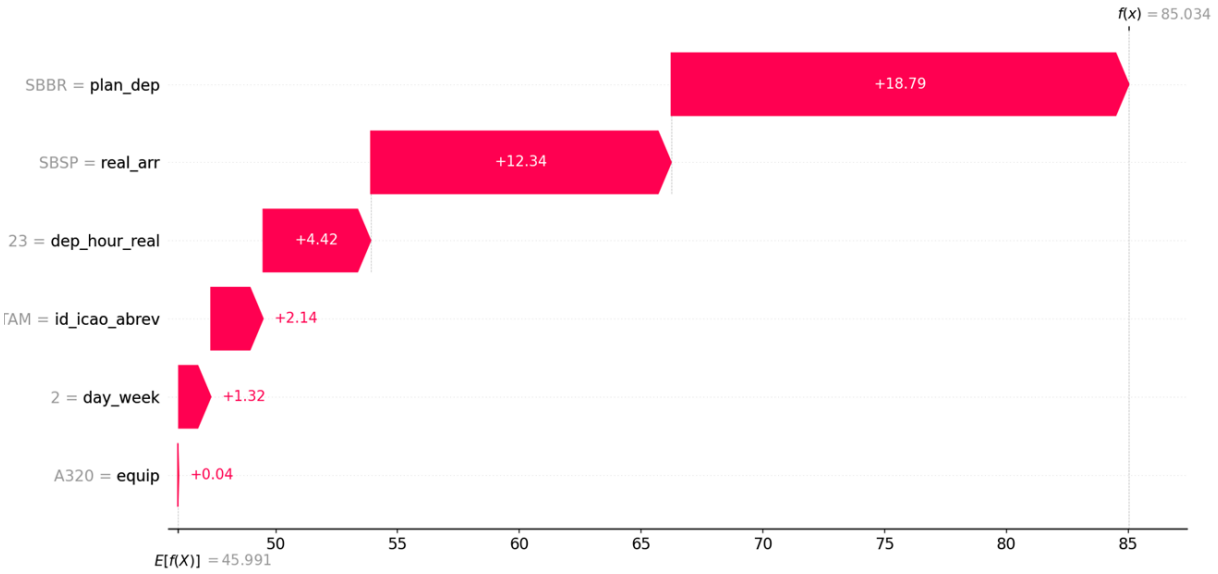


FIGURE 4.15 – SHAP explanations for an arrival flight at SBSP at 23:00.

TABLE 4.5 – Average SHAP value of the feature *hour* at different times of day for the selected airports.

Arrival Airport	**SHAP Value (08:00)**	**SHAP Value (23:00)**
SBGR	-2.14	7.94
SBSP	-4.72	4.42

Besides the local interpretation of the model's prediction using waterfall plots, an additional analysis of SHAP values is performed using boxplots that represent the distribution of the impact of each feature on model predictions at the 95% quantile. This type of plot is particularly useful for comparing dispersion, central tendency, and identifying the presence of possible outliers in the effects attributed by the model to each explanatory variable.

In the plots, each category within a feature (such as departure/arrival airport, hour, or aircraft type) is represented by a box that displays the interquartile range of SHAP values, while the horizontal line indicates the median of the distribution. As described above, positive SHAP values suggest a contribution to increasing the predicted trajectory deviation relative to the flight plan, whereas negative values indicate the opposite. Thus, the boxplots allow to observe how the impact of a variable varies across different groups or categories, highlighting distinct model behaviors under different operational conditions. These plots are shown in Figures 4.16, 4.17, 4.18, 4.19, 4.20, and 4.21.

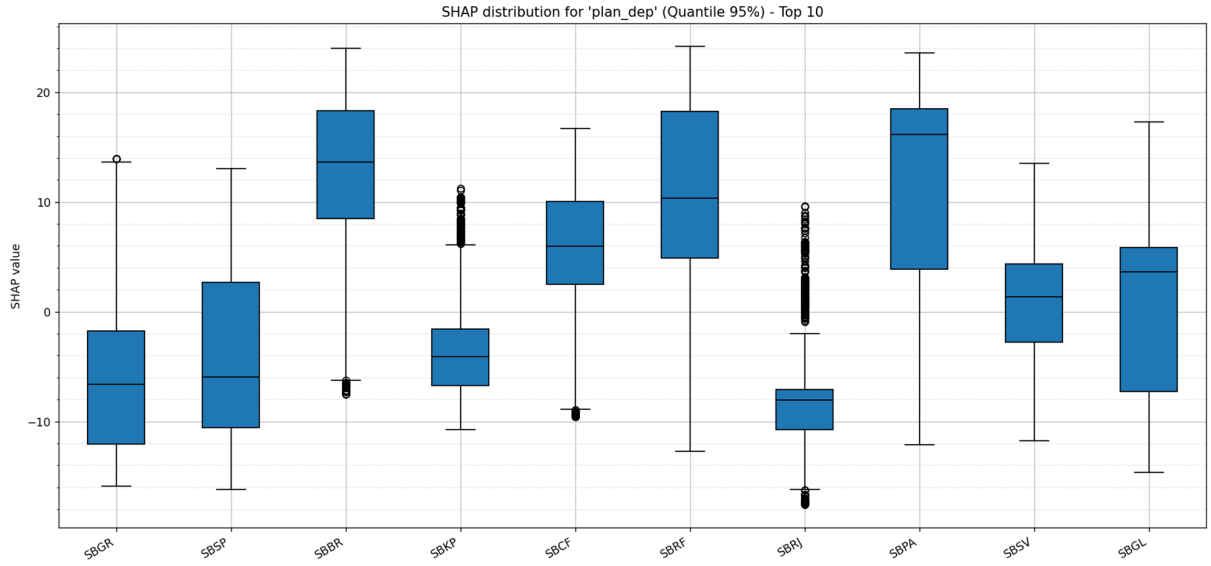


FIGURE 4.16 – SHAP values by departure airport (95% quantile).

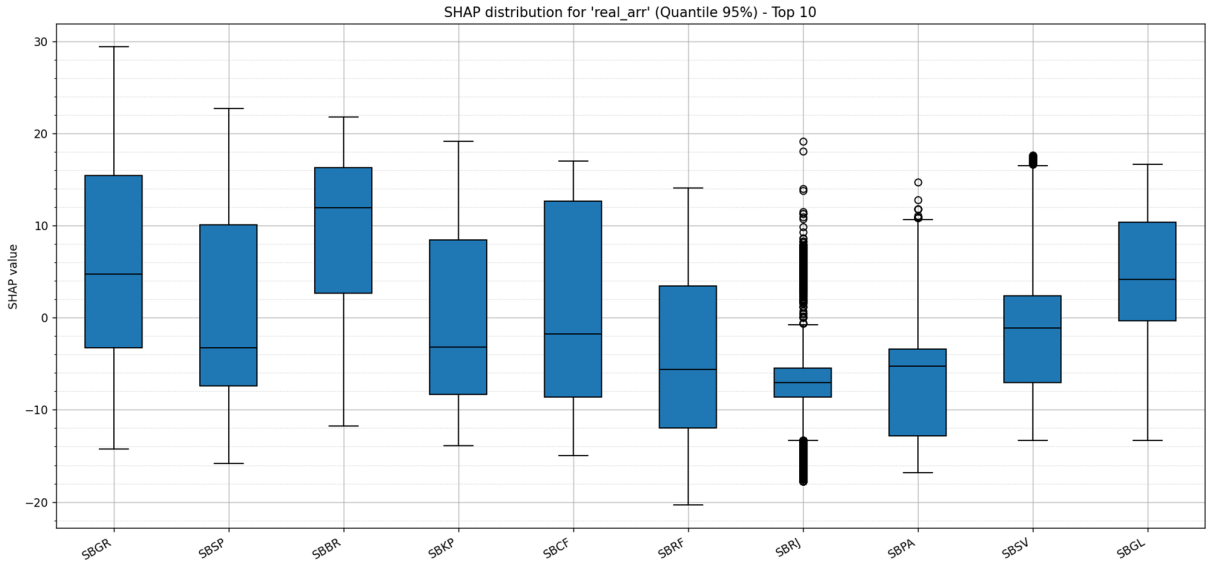


FIGURE 4.17 – SHAP values by arrival airport (95% quantile).

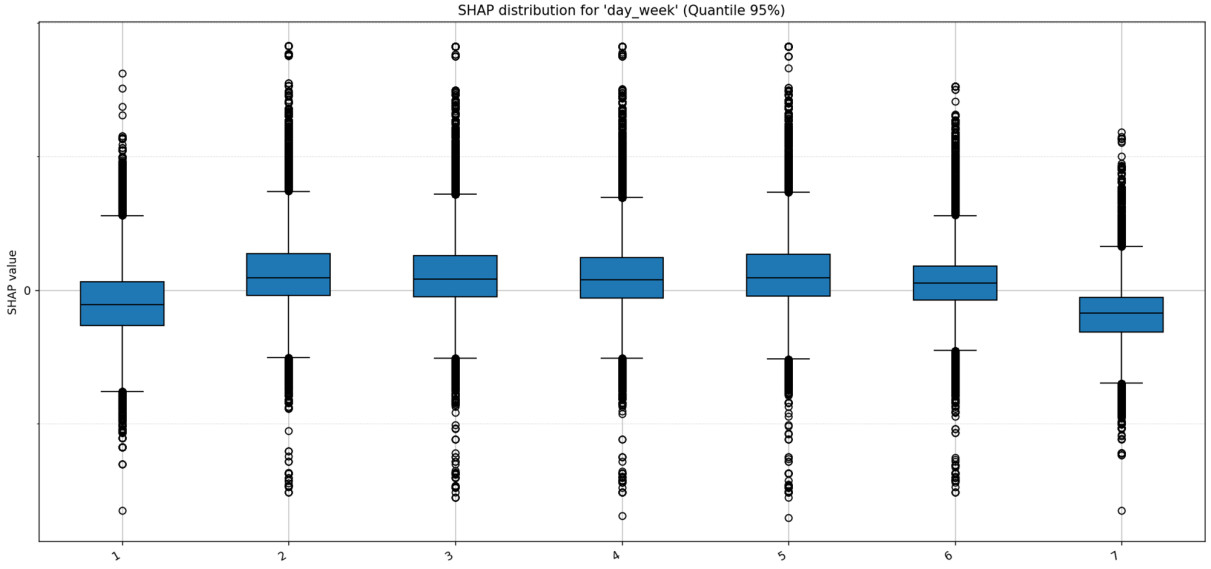


FIGURE 4.18 – SHAP values by day of the week (95% quantile).

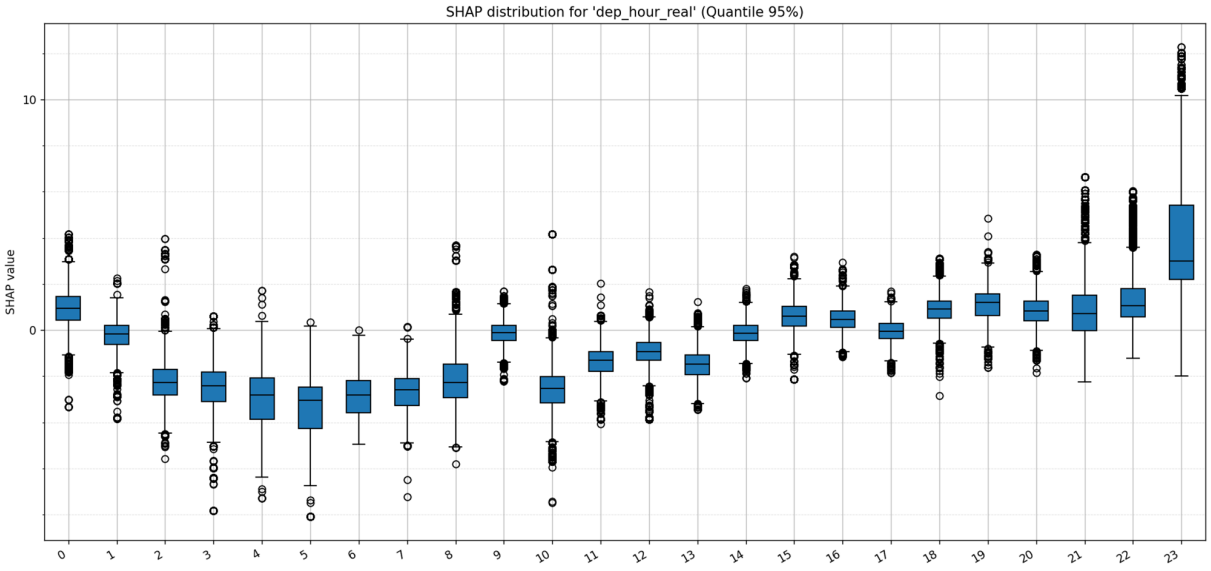


FIGURE 4.19 – SHAP values by departure hour (95% quantile).

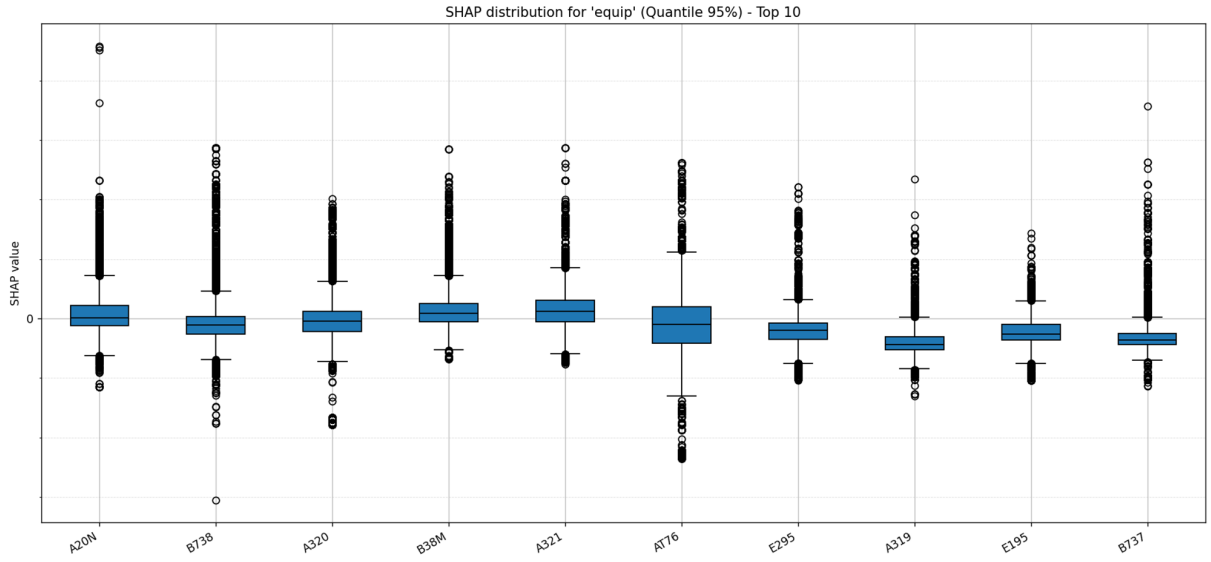


FIGURE 4.20 – SHAP values by aircraft type (95% quantile).

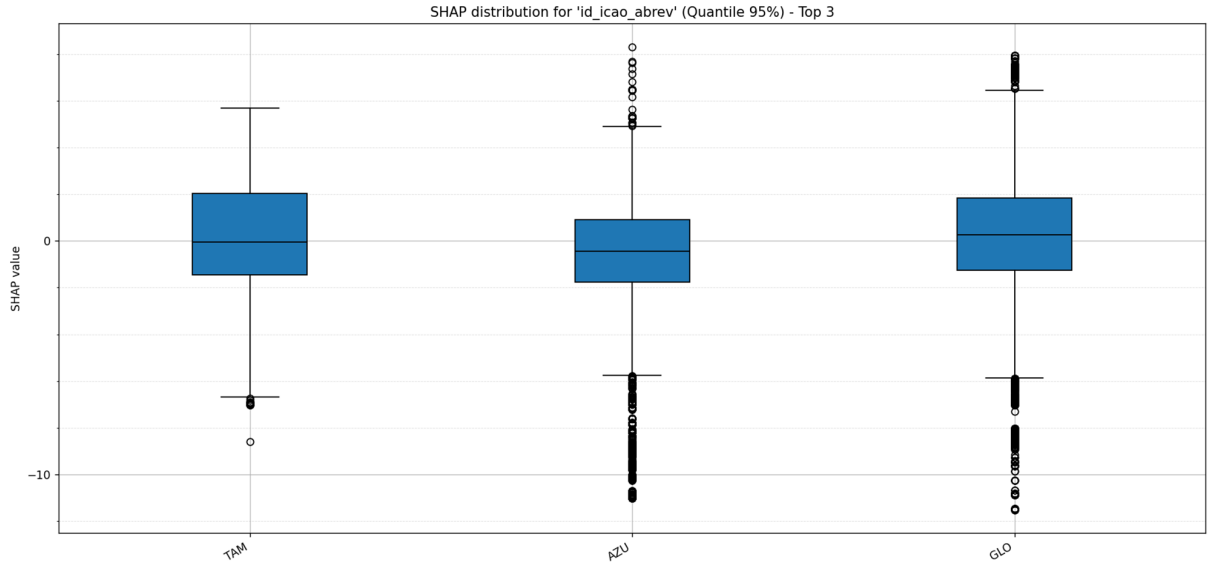


FIGURE 4.21 – SHAP values by airline (95% quantile).

The visualizations bring several insights. It is interesting to note the significant variability of SHAP values across departure and arrival airports. In Figure 4.16, SBBR, SBRF and SBPA show significantly higher SHAP values, suggesting that flights departing from these airports are more likely to experience larger trajectory deviations from the flight plan. In contrast, Figure 4.17 reveals that flights arriving at SBGR and SBBR tend to experience larger trajectory deviations. Figure 4.21 shows that the SHAP values for Azul are slightly lower, suggesting a more predictable en-route performance for flights performed by this airline. This might be associated with the use of less congested routes. Indeed, it is possible to note that a higher flow of aircraft tends to result in higher SHAP values, which, in turn, are associated with greater predicted trajectory deviations. This

behavior can be observed, for example, in Figure 4.19, together with the histogram in Figure 4.4, where it is noted that during peak hours the model predicts larger deviations, while in periods of lower air traffic, the predictions indicate smaller trajectory deviations.

5 Final Considerations

This section presents the main conclusions and suggestions for future work.

5.1 Conclusion

This study focused on predictive modeling of en-route operational performance within the Brazilian airspace. Using historical operational data and machine learning techniques, we developed a multi-quantile regression model to predict the difference between the actual flown distance during the en-route phase and the planned distance specified in the flight plan. A key feature of our approach lies in its ability to quantify predictive uncertainty and generate local explanations, thereby enhancing its decision support potential for airlines and air traffic management in areas such as fuel planning and traffic flow management.

We specifically use gradient-boosted decision trees (GBDT) with the CatBoost algorithm to learn the multi-quantile regression model together with Shapley Additive Explanations (SHAP) to obtain local explanations for the predictions. The approach is demonstrated with historical data for one year of operations in the Brazilian airspace, including aircraft surveillance data and flight plan data. Compared with baseline predictions derived from basic statistical principles, the predictions made with the proposed method are found to be more precise, reducing the multi-quantile error by 77%.

The combined use of GBDT and SHAP enabled not only the generation of probabilistic forecasts but also the analysis of the relative contribution of each feature to the formation of these estimates. In this regard, the study integrates statistical performance and interpretability, a relevant aspect for the adoption of machine learning solutions in aviation operational contexts.

Overall, the results showed that the model was able to predict en-route trajectory performance with good consistency across most of the analyzed quantiles, preserving the correct ordering of predictions and presenting reduced pinball error values, especially for the extreme quantiles. Furthermore, the coverage analysis indicated a reasonable adherence between the predicted bands and the observed values, reinforcing the potential

of the proposed approach to achieve the objectives of the work.

5.2 Future Work

Based on the limitations and assumptions discussed in Section 3.5, several opportunities for continuity and improvement can be explored. The first involves the incorporation of meteorological information, such as wind fields, turbulence forecasts, and convective weather indicators, in order to capture deviations caused by variable environmental conditions. This inclusion tends to enhance the accuracy of the predictions.

Another possibility consists of expanding the temporal coverage of the dataset. Data from different periods would allow the model to learn seasonal patterns and structural variations in demand, increasing its generalization capability for other scenarios within the Brazilian airspace.

Additionally, the inclusion of variables related to air traffic management, such as traffic flow management restrictions, could make it possible to account for other operational factors that can influence en-route trajectory deviations. Analyzing the representativeness of data across airports and operators can also help mitigate imbalance effects observed in the current model.

Finally, a promising avenue involves a study aimed at integrating the model with operational decision support tools for fuel planning and traffic flow management, contributing to a more proactive and data-driven decision-making.

Bibliography

CATAIA, M.; GALLO, F. Sistema de transporte aéreo flexível e integração do território brasileiro. **Revista GeoNordeste**, 2007.

DALMAU, R.; FALCO, P. D.; SPAK, M.; VARELA, J. D. R. Probabilistic pre-tactical arrival and departure flight delay prediction with quantile regression: A case study for geneva international airport using operational data. *In*: **15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023)**. **Proceedings** [...]. [*S.l.*]: FAA/Eurocontrol, 2023.

DEWEZ, F.; GUEDJ, B.; VANDEWALLE, V. From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning. **Data-Centric Engineering**, v. 1, p. e11, 2020. Available at: https://doi.org/10.48550/arXiv.2005.05286.

FRIEDMAN; ROBERT, T.; TIBSHIRANI HASTIE, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009. (Springer Series in Statistics).

HOFFMAN, H. **A New Way to Predict Probability Distributions**. 2023. Accessed: April 15, 2025. Available at: https://towardsdatascience.com/a-new-way-to-predict-probability-distributions-e7258349f464/.

JONES, K. T. **Cross-Validation for Time Series Data**. 2025. Accessed: Oct. 25, 2025. Available at: https://medium.com/%40kylejones_47003/cross-validation-for-time-series-data-51fd11c38e2b.

LIU, Y.; HANSEN, M.; BALL, M. O.; LOVELL, D. J. Causal analysis of flight en route inefficiency. **Transportation Research Part B: Methodological**, Elsevier, v. 151, p. 91–115, 2021. Available at: https://doi.org/10.1016/j.trb.2021.07.003.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B.; KATZ, R.; HIMMELFARB, J.; BANSAL, N.; LEE, S.-I. From local explanations to global understanding with explainable ai for trees. **Nature Machine Intelligence**, v. 2, p. 56–67, 2020. Available at: https://doi.org/10.1038/s42256-019-0138-9.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. Cambridge, MA: MIT Press, 2012. (Adaptive Computation and Machine Learning).

MURçA, M. C. R.; GUTERRES, M. X.; de Oliveira, M.; Tarelho Szenczuk, J. B.; SOUZA, W. S. S. Characterizing the brazilian airspace structure and air traffic

performance via trajectory data analytics. **Journal of Air Transport Management**, Elsevier, v. 85, p. 101798, 2020. Available at: https://doi.org/10.1016/j.jairtraman.2020.101798.

PERFORMANCE REVIEW COMMISSION. **Performance Review Report PRR 2018**. Brussels, Belgium, 2018.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. **CatBoost: unbiased boosting with categorical features**. 2019. Available at: https://doi.org/10.48550/arXiv.1706.09516.

THIAGARAJAN, B.; SRINIVASAN, L.; SHARMA, A. V.; SREEKANTHAN, D.; VIJAYARAGHAVAN, V. A machine learning approach for prediction of on-time performance of flights. *In*: **36th Digital Avionics Systems Conference (DASC)**. **Proceedings** [...]. IEEE/AIAA, 2017. Available at: https://doi.org/10.1109/DASC.2017.8102138.

TXAPARTEGI, A.; CAZCARRO, I. The economic and environmental impact of limiting air routes where there is a rail alternative: a case study of spain. **Regional Environmental Change**, v. 25, p. 37, 2025. Available at: https://doi.org/10.1007/s10113-025-02366-0.

ZHOU, Z.-H. **Ensemble Methods: Foundations and Algorithms**. Boca Raton, FL: Chapman & Hall/CRC, 2012. (Chapman & Hall/CRC Machine Learning & Pattern Recognition Series).

ZHU, X.; HONG, N.; HE, F.; LIN, Y.; LI, L.; FU, X. Predicting aircraft trajectory uncertainties for terminal airspace design evaluation. **Journal of Air Transport Management**, v. 113, p. 102473, 2023. Available at: https://doi.org/10.1016/j.jairtraman.2023.102473.

ZHU, X.; ZHANG, K.; ZHANG, Z.; TAN, L. Predicting flight trajectory in convective weather through boosted spatiotemporal deep learning ensemble. **Journal of Advanced Transportation**, v. 2024, p. 6400839, 2024. Available at: https://doi.org/10.1155/2024/6400839.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO Nº	4. Nº DE PÁGINAS
TC	18 de novembro de 2025	DCTA/ITA/TC-094/2025	54

5. TÍTULO E SUBTÍTULO:

Predictive Modeling of En-Route Operational Performance in the Brazilian Airspace

6. AUTOR(ES):

Diogo Longo Polo

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Machine Learning; Air Traffic Management; Operational Efficiency

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Controle do tráfego aéreo; Planejamento estratégico; Combustíveis; Árvores de decisões; Aprendizagem (inteligência artificial); Segurança operacional; Transportes.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia de Civil-Aeronáutica. Orientadora: Profª Drª Mayara Condé Rocha Murça. Publicado em 2025.

11. RESUMO:

Accurately predicting the actual performance of a flight during its en-route phase, particularly deviations from the planned flight path, is crucial for optimizing fuel planning and airspace utilization. This study addresses the predictive modeling of en-route operational performance within the Brazilian airspace using machine learning techniques. We develop a multi-quantile regression model to estimate the deviation between the actual flown distance and the planned distance during the en-route flight phase. The model, learned with the CatBoost algorithm based on gradient-boosted decision trees, provides probabilistic forecasts and quantifies predictive uncertainty. Local interpretability is achieved through Shapley Additive Explanations (SHAP), providing insights into the relative influence of explanatory features. Using one year of operational data comprising aircraft surveillance and flight plan information, the proposed method outperforms baseline statistical approaches, reducing the multi-quantile error by 77%. By integrating machine learning techniques that combine predictive accuracy with interpretability, the proposed approach aims to deliver valuable decision support for airlines and air traffic management, particularly in areas such as fuel planning and traffic flow management.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**