Flight Delay Prediction Using Explainable AI

Francisco Tavares[1,2], Rodrigo Batista[1,2] and Rodrigo Taveira[1,2]

[1]Faculty of Science of the University of Porto
[2]Faculty of Engineering of the University of Porto

Abstract

This report investigates the use of explainable artificial intelligence (XAI) techniques to interpret machine learning models for flight delay prediction using operational, weather, and geolocation data. A structured explainability workflow is adopted, covering pre-modelling, in-modelling, and post-modelling stages. A Decision Tree is used as a glass-box model, while a Random Forest, serving as the black-box model, is analysed through post-hoc XAI methods, including surrogate models, SHAP, LIME, counterfactual explanations, and nearest neighbors. Explanation quality is quantitatively evaluated using comprehensiveness and sufficiency metrics. Results indicate that features, such as Airline and Departure time play a dominant role in delay prediction; however, removing these features leads to limited degradation in prediction confidence, suggesting that explanations based on single feature groups are insufficient. Furthermore, the study demonstrates that individual XAI methods, such as LIME, provide partial insights, reinforcing the need for combining multiple explanation techniques to obtain reliable and meaningful interpretations of complex predictive models.

Introduction

Flight delays are a persistent challenge in the aviation industry, affecting passengers, airlines, and airport operations. Although machine learning models have shown promising results in delay prediction, predictive performance alone is often insufficient for real-world adoption. In operational settings, understanding the reasons behind model predictions is essential to support trust, validation, and decision-making, a need addressed by Explainable Artificial Intelligence (XAI).

In this work, we study a binary classification task for flight delay prediction using structured tabular data that combines operational flight information with weather and geolocation features. This allows the analysis of heterogeneous factors such as airline, departure time period, aircraft characteristics, and meteorological conditions, providing a realistic testbed for XAI techniques applied to tabular models.

The objective of this project is to compare how different XAI techniques explain model behaviour across multiple stages of the modelling pipeline. Pre-modelling explanations are explored through exploratory analysis, in-modelling explanations are obtained using a glass-box Decision Tree, and post-hoc explanation methods are applied to a Random Forest model. Rather than focusing on methodological details, we emphasise a comparative analysis of the insights and limitations of each approach in a real-world prediction scenario.

Task 1: Pre-Modelling Explanations

Before training any predictive model, we applied pre-modelling explainability techniques to explore the structure of the flight delay dataset, identify dominant patterns, and understand how different groups of variables relate to delay outcomes. The objective of this stage is to support data understanding and guide modelling choices, rather than to optimise predictive performance.

Exploratory Analysis

We begin by analysing the characteristics of the *US_flights_2023* dataset, which contains approximately six million flight records. This large scale introduces high computational costs and substantial variability. In addition, the target variable presents a moderate class imbalance, with a higher proportion of on-time flights. To improve feasibility and interpretability, we restrict our analysis to flights departing from Hartsfield–

Jackson Atlanta International Airport (ATL) and Denver International Airport (DEN), the two most frequent airports in the dataset. This selection significantly reduces dataset complexity while also yielding a more balanced target distribution (**55.6%** on-time flights and **44.4%** delayed flights).

To reduce feature redundancy and improve interpretability, several feature transformations were applied. Geographic coordinates, namely latitude and longitude, were replaced by a single Haversine distance feature, capturing the great-circle distance between origin and destination. The original date variable was decomposed into day and month components, while the year was discarded, as the dataset covers a single calendar year.

Regarding the weather variables, exploratory analysis showed no clear monotonic relationship between precipitation intensity and flight delays, as seen in Figure 1. However, a noticeable difference was observed between rainy and non-rainy conditions. Based on this observation, the continuous precipitation variable was transformed into a binary rain indicator, indicating whether rainfall was present at the airport.

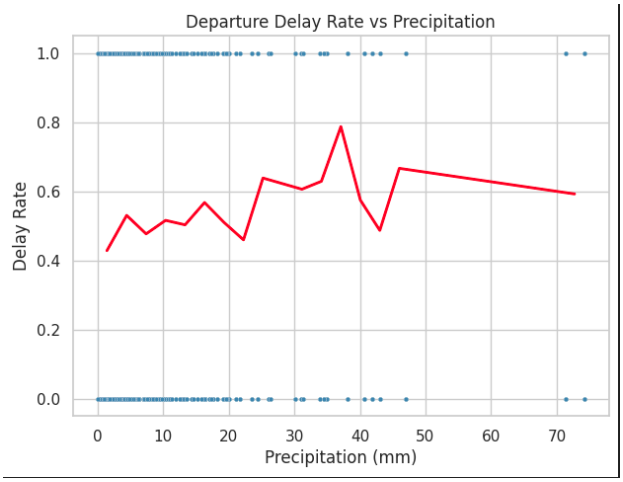

Figure 1: *Delay vs Precipation*

After removing redundant informative features, such as minimum and maximum temperature (given the presence of the average), as well as variables prone to data leakage, such as *Delay_Weather*, the remaining categorical features were encoded using Label Encoding. This choice avoids the dimensionality increase associated with one-hot encoding and is suitable for tree-based models, which do not rely on ordinal assumptions in encoded values.

Correlation heatmaps (Figure 2) revealed expected relationships among time-related variables, while categorical frequency plots for airline, time of day, and day of week exposed clear operational patterns. In particular, certain carriers and late departure periods consistently exhibited higher delay rates.

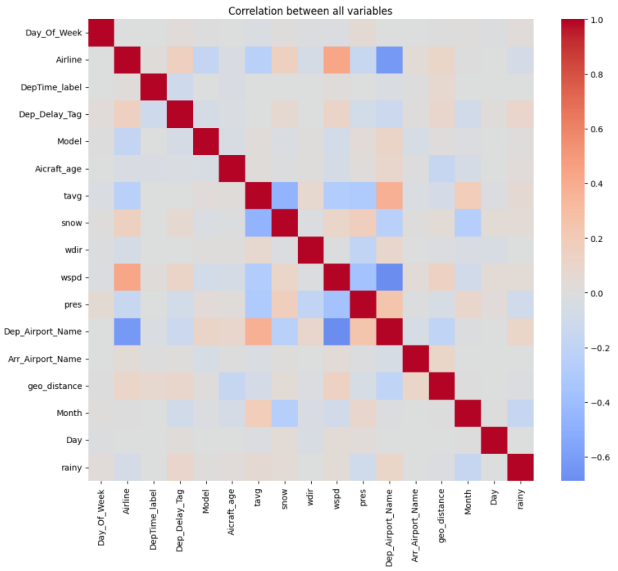

Figure 2: *Correlation betweeen variables*

Dimensionality Reduction

To investigate the global structure of the dataset, we applied dimensionality reduction techniques to the standardized numerical features.

PCA Principal Component Analysis (PCA) was first used to assess variance distribution across components. Projections onto the first two principal components, Figure 3, exhibit substantial overlap between delayed and on-time flights, indicating that simple linear combinations of numerical features do not provide clear class separation.

t-SNE To further explore potential non-linear structure, we applied t-SNE to a randomly 50000 sampled subset of the data, selected to balance computational feasibility and representativeness. The resulting embedding, Figure 4, shows strong overlap between delayed and non-delayed flights, with no clearly separable clusters associated with either class, even though

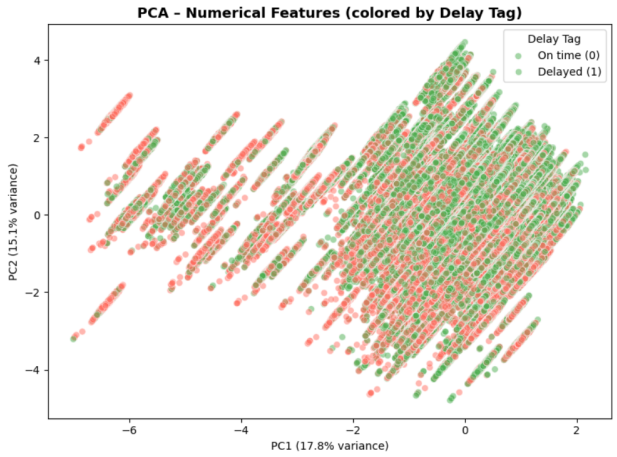

Figure 3: *PCA*

some local groupings are visible. This again, suggests that delay outcomes are driven by complex, context-dependent interactions rather than by easily separable low-dimensional patterns.

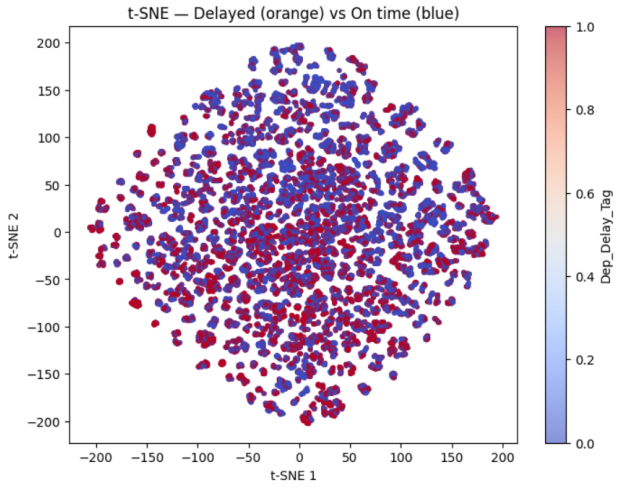

Figure 4: *t-sne*

Task 2: In-Modelling Explanations

To analyse model interpretability during training, we adopt an in-modelling explainability approach by training an inherently interpretable Decision Tree classifier. This model provides transparent decision logic, explicit feature usage, and human-readable rules, making it suitable for understanding how predictions are formed.

Glass-Box Model and Setup

Since the primary objective is explainability rather than optimal predictive performance, the tree depth was constrained to ***max_depth* = 6** and trained using a standard 80/20 train–test split. This limits model complexity while still allowing non-linear splits and basic feature interactions.

Predictive Performance

The Decision Tree achieved an overall accuracy of approximately **0.65**. For on-time flights (class 0), the model attained a high recall of **0.82**, correctly identifying most non-delayed flights, in contrast to a recall of **0.44** for delayed flights (class 1).

Global feature importance analysis indicates that the model relies heavily on a small set of variables, particularly *Airline*, *DepTime_label*, and *Month*, while other variables—such as aircraft model, weather indicators, flight distance, and temperature—exhibit noticeably smaller contributions. This concentration of importance reflects the model's preference for simple, high-level operational patterns.

By directly inspecting the learned decision rules, in Figure 5, it is evident that the top splits in the tree are primarily based on airline, time-of-day, and seasonal variables. This suggests that delay risk is explained through structured and interpretable rules. For instance, certain airlines are associated with higher delay likelihood, later departure periods are more delay-prone, and seasonal effects modulate overall risk.

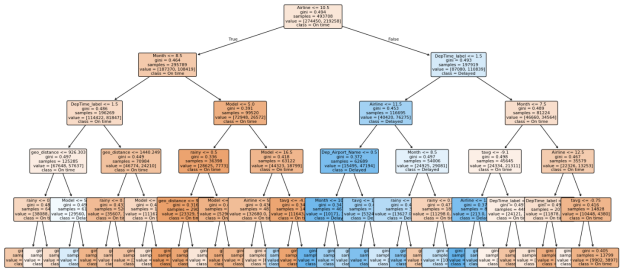

Figure 5: *decision tree*

Task 3: Post-Modelling Explanations

While glass-box models offer direct interpretability, their limited expressive power can restrict predictive performance in complex domains

such as flight delay prediction. To better capture non-linear relationships, we trained a Random Forest classifier as a black-box model and used it as the basis for post-hoc explainability analysis.

The Random Forest was trained with 200 trees, a maximum depth of 30, and regularization through minimum split and leaf sizes. It achieved an overall accuracy of approximately **0.70**, improving upon the performance of the decision tree. Nevertheless, a substantial number of delayed flights remain misclassified as on-time, highlighting the intrinsic difficulty of this dataset.

Given the reduced transparency of the Random Forest, post-hoc XAI techniques are required to interpret its decision-making process. Accordingly, we apply explanation methods from three complementary families: simplification-based, feature-based, and example-based approaches, which are analysed in the following sections.

Task 3.1: Simplification-Based Technique

To obtain a global and human-interpretable approximation of the Random Forest, we trained a surrogate decision tree on the predictions of the black-box model rather than on the ground-truth labels.

A surrogate tree with a maximum depth of four was selected as a compromise between interpretability and fidelity. The surrogate achieved a fidelity score of **0.792**, indicating that it captures a substantial portion of the Random Forest's decision logic. Additional experiments with tree depths ranging from two to eight confirmed the expected trade-off: increasing the depth improves fidelity to the black-box model but leads to more complex and less interpretable trees.

In terms of predictive performance, the surrogate decision tree reached **0.64**. This reduction is expected, as the surrogate is intentionally constrained to remain interpretable and does not aim to match the black-box model's predictive power.

Overall, these results illustrate the typical trade-off inherent in simplification-based XAI techniques: surrogate models sacrifice predictive accuracy in exchange for transparency, yet can still provide a meaningful and interpretable global summary of a complex model's decision process.

Task 3.2: Feature-Based Techniques

Feature-based methods aim to quantify how individual variables influence predictions. We applied SHAP and LIME, two widely used but conceptually different approaches, to compare the insights they provide.

SHAP SHAP explanations were computed using TreeExplainer to analyse feature contributions to the Random Forest predictions. Due to the high computational cost of SHAP for ensemble models, the global analysis was performed on a random subset of 500 test instances.

At the global level, SHAP summary plots, Figure 6, show that operational and temporal variables dominate the model behaviour, with Airline, DepTime_label, and Month having the largest impact on predictions. Weather-related features, such as wind speed and precipitation indicators, also contribute, but their influence is generally secondary. These results are consistent with the patterns observed during the pre-modelling analysis.

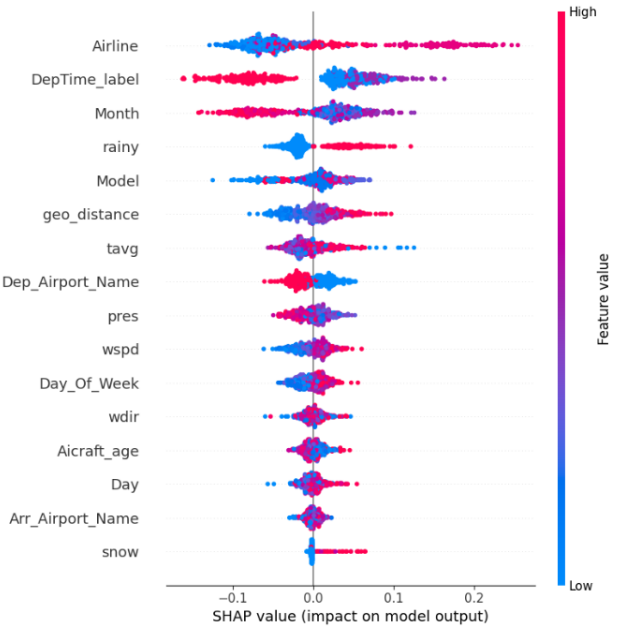

Figure 6: *Global SHAP*

A local explanation, as seen in Figure 7, was analysed using a SHAP waterfall plot for a correctly classified delayed flight (instance 3). Start-

ing from the baseline prediction (**E[f(X)] = 0.444**), the model output (**f(x) = 0.562**) exceeds the decision threshold. The strongest positive contributions arise from DepTime_label, geo_distance, and Month, with wind speed also increasing delay probability. In contrast, rainy, average temperature, and Airline slightly reduce the predicted delay probability, though not enough to offset the dominant positive effects. Overall, SHAP provides consistent and additive explanations that faithfully reflect the Random Forest's decision process at both global and local levels.

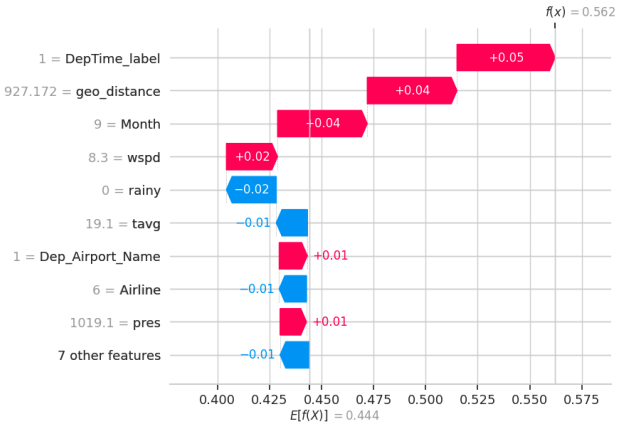

Figure 7: *Local SHAP (instance 3)*

LIME LIME was also applied to generate a local explanation by fitting linear surrogate models around a individual prediction of the Random Forest. For the analysed delayed instance (instance 3), LIME assigned a high probability (**0.68**) to the delayed class, Figure 8. The explanation indicates that DepTime_label, aircraft model, Month, Airline, and geo_distance are the main features pushing the prediction toward delay, while weather-related variables such as rainy, snow, and average temperature slightly reduce the predicted delay probability. This local explanation is consistent with the SHAP analysis, reinforcing the dominance of operational and temporal factors in the model's decision.

To assess explanation reliability, we evaluated the local fidelity of LIME reaching an ($\mathbf{R^2}$ **of 0.19**). This indicates that the linear surrogates often fail to approximate the complex, non-linear decision boundary of the Random Forest. While this limits LIME's faithfulness in such settings, it remains useful as a qualitative interpretability tool, providing intuitive insights into the main features influencing individual predictions rather than an exact local representation of the model.

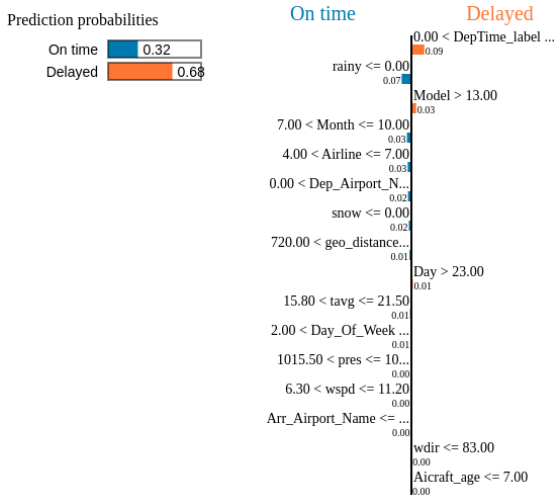

Figure 8: *LIME*

Task 3.3: Example-Based Technique

Counterfactual explanations were generated using DiCE to analyse how realistic changes to input features could alter the model's predictions. For a flight originally predicted as delayed (class 1), Figure 9, valid counterfactuals were obtained by modifying operational variables such as the departure time-of-day and aircraft model, together with changes in average temperature and wind-related conditions. This counterfactual satisfies validity and feasibility, and is partially actionable, since scheduling-related variables can be influenced, whereas weather conditions cannot be directly controlled.

Conversely, for a flight initially predicted as on-time (class 0), producing a delayed prediction required changes in the operating airline and aircraft model, combined with environmental conditions such as temperature values and wind direction. While the generated counterfactual is valid, its actionability is also limited, as airline assignment and weather variables are largely outside user control. Nevertheless, the explanation remains plausible, as the suggested feature values are consistent with realistic operational and meteorological scenarios. A curious observation is that, across all generated counterfactuals, the wind direction feature is set to zero. This suggests that the model either assigns

limited importance to this variable or that setting it to a neutral value is an efficient way for satisfying the desired prediction change.

Counterfactual to flip Delayed -> On time (predicted 1 -> 0)

	original	counterfactual
Day_Of_Week	4.000000	4.0000
Airline	6.000000	6.0000
DepTime_label	1.000000	2.0000
Model	14.000000	1.0000
Aicraft_age	1.000000	6.0000
tavg	19.100000	7.0000
snow	0.000000	14.2000
wdir	58.000000	0.0000
wspd	8.300000	8.3000
pres	1019.100000	1019.1000
Dep_Airport_Name	1.000000	1.0000
Arr_Airport_Name	10.000000	10.0000
geo_distance	927.171778	927.1718
Month	9.000000	9.0000
Day	28.000000	28.0000
rainy	0.000000	0.0000

Figure 9: *Counterfactual*

Nearest Neighbors

Nearest neighbors analysis was used to provide example-based explanations by comparing each instance with similar flights in the dataset. For the instance predicted as delayed, most of the closest neighbors correspond to the same route and share highly similar operational characteristics, with delayed outcomes persisting even across an extended neighborhood. This suggests that the route itself is structurally prone to delays. More details about this method are present on the notebook.

Task 4: Quality of Explanations

To quantitatively assess explanation quality, we evaluated SHAP-based explanations using functionally grounded metrics that rely solely on model outputs. In particular, we analysed comprehensiveness and sufficiency, which measure whether the features identified by the explanation are necessary and/or sufficient for a given prediction. Both metrics were computed for a representative test instance using the top five features ranked by absolute SHAP value.

Comprehensiveness

Comprehensiveness measures the change in the model's confidence when the most important features are removed from the input. For the analysed instance, the predicted probability increases from **0.562** to **0.624**, resulting in a comprehensiveness score of **-0.062**. This relatively small reduction indicates that, while the selected features contribute to the prediction, removing them does not drastically alter the model's output, suggesting that the model distributes information across multiple features.

Sufficiency

Sufficiency evaluates whether the most important features alone are sufficient to reproduce the original prediction when all other features are masked. Using only the top five features leads to a prediction of **0.393**, compared to **0.562** for the full model. This substantial difference shows that the selected features are not sufficient on their own to explain the model's decision and that additional contextual variables play an important role.

Conclusions

This study compared multiple explainable AI techniques for flight delay prediction using a structured pre-modelling, in-modelling, and post-modelling workflow. Pre-modelling analysis revealed that delays arise from complex, non-linear interactions between operational, temporal, and weather-related factors. An interpretable Decision Tree provided transparent but limited explanations, while a Random Forest improved performance at the cost of interpretability. Post-hoc methods showed complementary strengths: SHAP offered stable explanations, LIME provided intuitive but less faithful local insights, and surrogate and counterfactual approaches added global and actionable perspectives. Overall, the results confirm that no single XAI method is sufficient, and that combining multiple approaches is necessary to reliably interpret complex models.