OPEN IIT
DATA ANALYTICS

D-35

INDEX

INTRODUCTION

In today's competitive aviation industry, airlines face the critical challenge of enhancing profitability while managing a multitude of operational areas, including flight scheduling, ground crew management, fuel optimisation, and passenger services. Due to strict regulatory requirements, it is essential for airlines to optimise these processes to minimise delays, improve turnaround times, and increase overall operational efficiency.

Problem Overview

The focus of this report is to develop a predictive solution that supports data-driven decision-making across multiple facets of airline operations. Specifically, the report addresses two core objectives:

Flight Delay Prediction: By analysing historical data, the model seeks to predict potential delays in scheduled flight operations. This predictive capability enables airlines to proactively allocate resources, reducing the impact of delays.

Flight Rescheduling: The model aims to provide an optimised rescheduling solution for flights, minimising cumulative delays and improving resource management.

This solution is designed to be scalable and adaptable across various airports, aircraft, and routes, making it suitable for the real-world complexities inherent in airline operations.

OVERVIEW OF THE DATASET

To develop an effective model for predicting arrival flight delays, we needed a comprehensive dataset that captured various factors influencing delays across a broad range of flights. Since no specific dataset was provided, we utilised web scraping techniques to compile necessary data from public sources.

Our primary data source was the Bureau of Transportation Statistics (BTS) website, transtats.bts.gov, which provides detailed flight data for the United States. We collected flight records for the year 2023, initially yielding a dataset of approximately 6.6 million records. Due to the large volume, processing this entire dataset was computationally prohibitive.

To streamline the dataset, we focused on flights from the top five busiest airports in the United States:

- Hartsfield-Jackson Atlanta International Airport (ATL)
- Denver International Airport (DEN)
- Charlotte Douglas International Airport (CLT)
- Chicago O'Hare International Airport (ORD)
- Dallas Fort Worth International Airport (DFW)

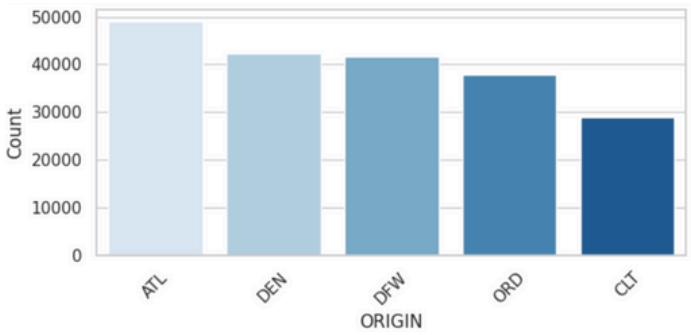

Fig-1: Count of data from each airport

This initial filtering reduced the dataset to 1.3 million records, making it more manageable. To further enhance computational efficiency, we then proportionally downsampled the dataset to approximately 200,000 rows. This sampling was designed to retain the statistical distribution and essential characteristics of the original data while making model training and testing feasible.

EXPLORATORY DATA ANALYSIS

Through exploratory data analysis, we gained several key insights into the structure and characteristics of the dataset. Visualizations facilitated a deeper understanding of feature distributions, relationships, and trends. Key insights include:

Monthly Distribution: Analysis showed an even distribution of flights across all months, indicating that seasonality effects are consistent and balanced throughout the dataset. (refer Fig 2)

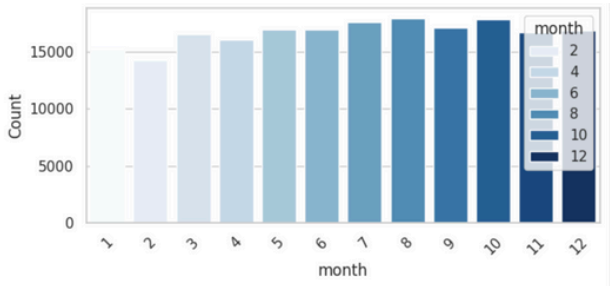

Fig-2: Monthly count plot

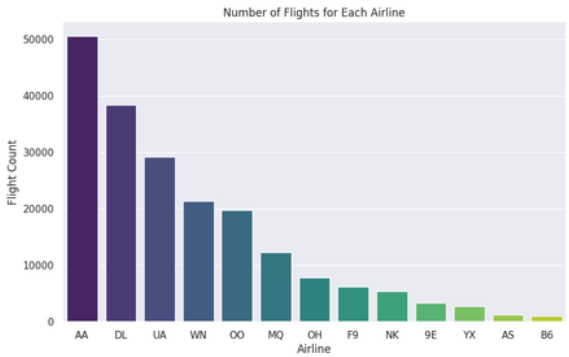

Fig-3: Flight Frequency per airport

Airline Representation: The data reveals that the top three airlines - American Airlines (AA), Delta Air Lines (DL), and United Airlines (UA) - account for nearly half of the dataset entries. Although a total of 14 airlines are represented, these three airlines dominate flight frequency (Figure 3 shows the Number of Flights per Airline).

Hourly Delay Patterns: The occurrence of delays was well-distributed across various hours of the day. However, flights scheduled between 12 A.M. and 5 A.M. rarely experienced delays, likely due to reduced airport activity during these hours (as shown in Figure 4, Delay Distribution by time group).

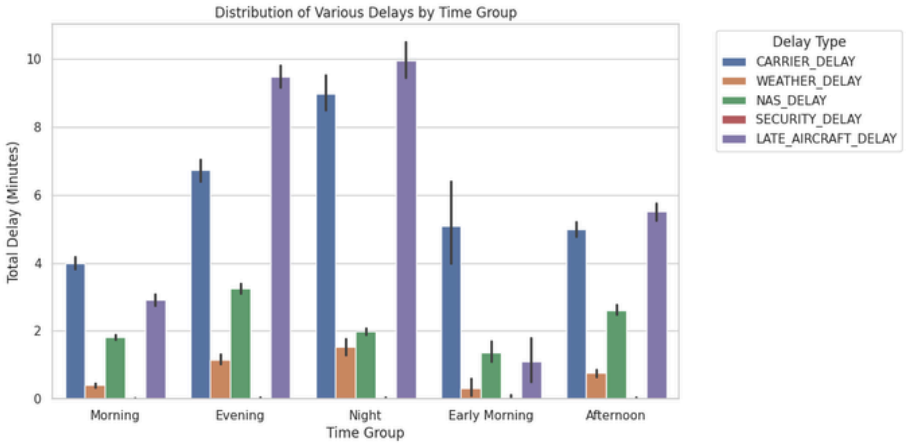

Fig-4: Delay distribution by time group

Delay Frequency: Approximately 60% of flights in the dataset have no recorded delay, with either early or on-time arrivals. For the remaining records, delay duration follows a decreasing distribution from 0 to 500 minutes, as visualized in Figure 5. This distribution highlights that extreme delays are rare, and most delays fall within a lower range.

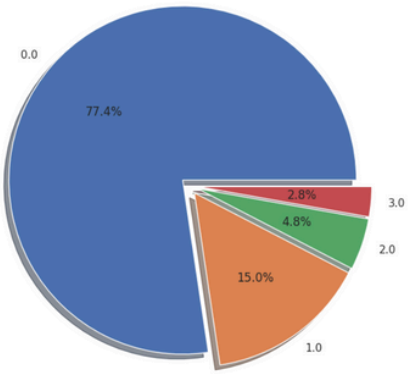

Fig-5: Delay status

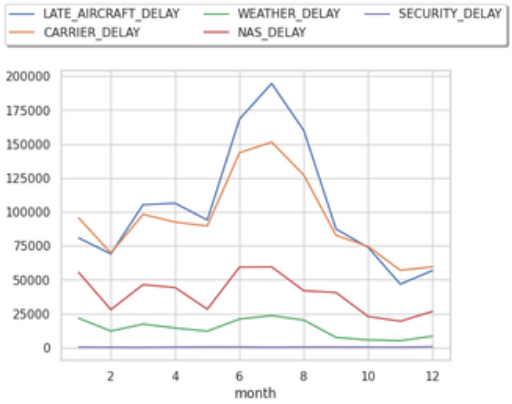

Fig-6: Monthly delayed flights

Seasonal Delay Patterns: Delay factors such as weather, security issues, and carrier-specific delays vary noticeably across different months. Delays due to these factors increase in specific months, aligning with seasonal weather changes and travel patterns (Figure 6 shows the Monthly Delay Range).

Carrier-Specific Delays: Certain airlines, notably B6 and F9, exhibit high levels of carrier-related delays with significant outliers, as indicated in the data. This finding confirms the effectiveness of the categorical feature Carrier Delay in capturing airline-specific delay trends (see Figure 7, Carrier Delay Distribution per Airline).

Distance vs. Delay Correlation: Analysis indicates a positive correlation between flight distance and delay, with longer flights more prone to higher delays. This pattern, while intuitive, is reinforced by the data and provides a useful predictor of delays (illustrated in Figure 8, Distance vs. Delay Plot).

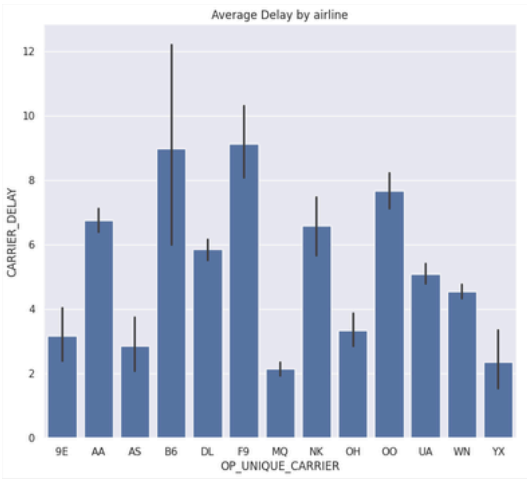

Fig-7: Carrier delay distribution per airline

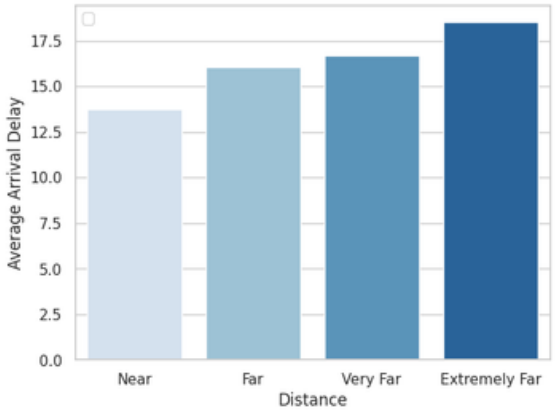

Fig-8: Distance vs delay

DATA PREPROCESSING

The dataset comprises multiple categorical columns and several numerical columns, each requiring distinct preprocessing methods to ensure compatibility with machine learning models. Below, we outline the specific techniques applied to handle these data types.

The numerical columns in the dataset displayed significant variation, with many values unnormalized and numerous outliers. Consequently, we applied the Robust Scaler, which scales data based on the interquartile range, making it well-suited for data containing outliers. This scaling improved the consistency of the numerical data without being affected by extreme values.

For categorical features, we employed One-Hot Encoding and Label Encoding to convert categorical data into numerical form. This transformation allows categorical features to be used with various machine learning models. One-Hot Encoding was used for columns with nominal categories, whereas Label Encoding was applied to ordinal categories where order is implied.

OUTLIER REMOVAL

To enhance the quality of the dataset, we employed HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) as an outlier removal technique. This method is particularly effective for identifying outliers in large datasets, as it can detect clusters of varying densities and classify points that do not belong to any cluster as noise.

In our analysis, HDBSCAN successfully identified approximately 29,000 data points as outliers. These outliers were subsequently removed from the dataset to improve the robustness of the model.

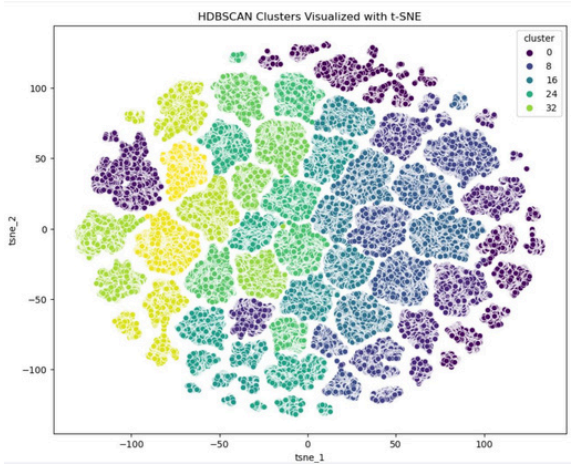

Fig-9: Visualisation of Clusters

FEATURE ENGINEERING AND SELECTION

Using the filtered dataset, we developed several additional features known to influence flight delays and also replaced few features that are present in the raw scrapped dataset but realistically won't be present when prediction is needed,These engineered features included:

- Congestion: Calculated as the daily number of flights per airport per airline, offering insights into delay patterns associated with high airport activity.
- Peak Hours: Binary indicators for high-traffic times, capturing periods when delays are more likely.
- Weather-Based Features: Factors like temperature, humidity, and wind speed have been scrapped from an API Meteostat.
- Average Carrier Delay: The weekly average delay per airline, allowing the model to account for historical delay tendencies of specific carriers.

These new features, combined with the downsized dataset, provided a strong foundation for training and evaluating our model effectively.

The dataset includes both categorical and numerical features. Categorical features, such as 'OP_UNIQUE_CARRIER' and 'ORIGIN', are explicitly cast as strings, which allows CatBoost to recognize them as categorical during model training. The target variable is arrival delay (ARR_DELAY), and cluster is excluded from the feature set. Data is split into training and testing sets, with 25% of the data reserved for testing.

Feature Selection with CatBoost

The Pool objects are created to encapsulate the training and testing data, ensuring that categorical features are correctly handled. A CatBoostRegressor model is initialized with 1000 iterations and a fixed random seed for reproducibility. The SHAP-based feature selection is performed using the select_features method. This method evaluates the importance of features based on their SHAP values, indicating their contribution to predictions. The selection algorithm employed is RecursiveByShapValues, which iteratively removes less important features, retaining only the top 20 features from the first 26 evaluated.

The final model is trained using the selected features, and visualizations are generated to illustrate the feature selection process.

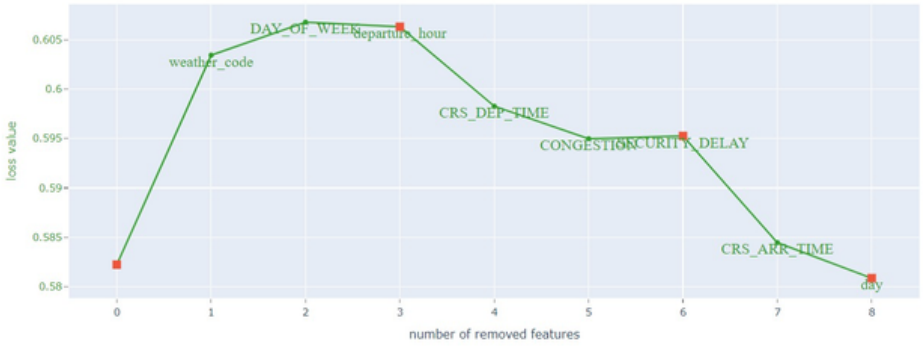

Fig-10: Feature removal

Selected Features

The following columns are retained after evaluation and feature selection

- 'QUARTER'
- 'month'
- 'ORIGIN'
- 'DAY_OF_WEEK'
- 'DEP_DELAY'
- 'TAXI_OUT'
- 'CRS_ELAPSED_TIME'
- 'DISTANCE'
- 'SECURITY_DELAY'
- 'LATE_AIRCRAFT_DELAY'
- 'weekly_carrier_delay'
- 'departure_hour'
- 'peak_1'
- 'peak_2'
- 'temp(in celsius)'
- 'wind_speed'
- 'dew_point'
- 'humidity'
- 'pressure'
- 'weather_code'
- 'precipitation'
- 'OP_UNIQUE_CARRIER',

MODEL TRAINING AND EVALUATION

In the analysis of flight delay predictions, various machine learning models were employed to determine their effectiveness in capturing the underlying patterns in the data. The models evaluated included CatBoost, XGBoost, Linear Regression, and LightGBM (LGBM). Each model's performance was measured using the R^2 score, a statistical metric that represents the proportion of variance for the dependent variable that is explained by the independent variables in the model. An R^2 score closer to 1 indicates a better fit of the model to the data.

The R^2 scores for the models were as follows:

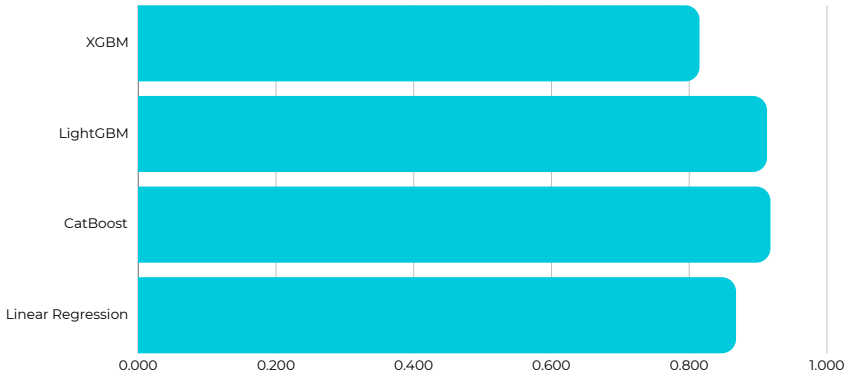

Fig-11: R2 score of different models

CatBoost: 0.9183, XGBoost: 0.8153, Linear Regression : 0.8681, LightGBM: 0.9134

From the results, CatBoost outperformed the other models, showcasing its ability to accurately predict flight delays. This performance indicates that CatBoost effectively captures the complexities of the dataset, providing reliable predictions.

Hyperparameter Tuning

To enhance the performance of the CatBoost model, a systematic hyperparameter tuning process was conducted. Hyperparameter tuning involves adjusting the parameters of a machine learning model to optimize its performance. The tuning process is critical because the default settings may not yield the best results for specific datasets.

Through this tuning process, the following optimal hyperparameters were identified: Bootstrap Type: Bernoulli, Depth: 7, Iterations: 977, L2 Leaf Regularization: 1, Learning Rate: 0.1536

These parameters were selected based on their impact on the model's ability to generalize and perform well on unseen data. After implementing these tuned parameters, the CatBoost model demonstrated a remarkable improvement in its performance, achieving a post-tuning R^2 score of 0.9248. This signifies a substantial increase in the model's explanatory power regarding the variance in flight delay data.

Additionally, the Root Mean Squared Error (RMSE) for the post-tuning CatBoost model was calculated to be 0.48, indicating the average prediction error. A lower RMSE value signifies that the model's predictions are closer to the actual values, further confirming the effectiveness of the tuned CatBoost model.

EXPLAINABILITY

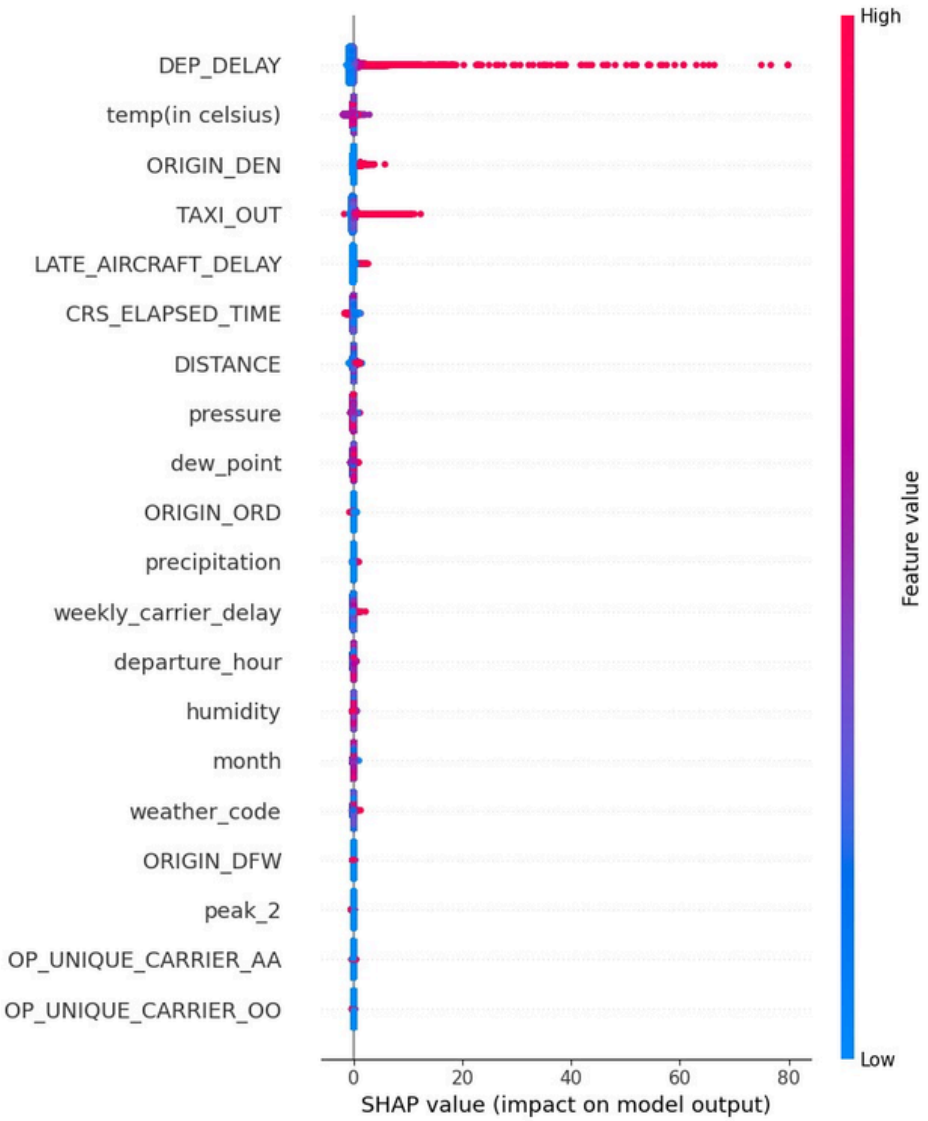

Fig-12: SHAP to explain feature impact on model

FLIGHT RESCHEDULING

In addressing the complex issue of flight rescheduling, we employed the Gurobi optimization library, renowned for its efficiency in tackling intricate optimization challenges. The primary constraints we faced included airport capacity limitations, allowable delays, and the imperative to minimize passenger disruptions. Consequently, we developed a robust optimization framework designed to dynamically reschedule flights, ensuring efficiency in both arrival and departure times while adhering to operational limits.

Gurobi Optimization

Gurobi Optimization serves as a cutting-edge solver tailored for resolving complex optimization problems. Its capabilities encompass a wide array of mathematical programming tasks, including Linear Programming (LP), Quadratic Programming (QP), and Mixed-Integer Programming (MIP). The latter is particularly valuable in the context of flight scheduling, where rescheduling decisions necessitate the use of integer and binary variables to represent constraints and flight slot assignments.

Using airline data from January 2023 at Atlanta Airport, we formulated a Gurobi model designated "Flight_Rescheduling." Within this model, we introduced decision variables that represented the necessary adjustments for rescheduling each flight. Specifically, these included positive and negative adjustments for both arrival and departure flights. Positive adjustments indicate the extent to which a flight is rescheduled later, whereas negative adjustments denote how much it is rescheduled earlier. By incorporating these displacement variables, we aimed to effectively address scheduling challenges while minimizing disruptions to the overall flight timetable. To facilitate scheduling within the constraints of a 24-hour operational period, we established 288 discrete time slots, each corresponding to 5-minute intervals throughout the day. The binary decision variables were utilized to indicate whether a flight would be scheduled in a specific time slot, thereby capturing the scheduling dynamics and enabling us to frame the problem as an optimization task.

Objective Function

The objective function is a crucial component guiding the optimization process. Our primary goal was to minimize the total displacements of flights while ensuring maximum resource utilization to minimize delays. To achieve this, we constructed the objective function by summing various components.

We aggregated the positive and negative displacement variables for all arrival flights, represented as u_plus and u_minus. This summation reflects the total amount of rescheduling necessary for each flight arriving at Atlanta Airport (ATL). A similar aggregation was performed for departure flights with the corresponding variables v_plus and v_minus. Furthermore, we included a term that quantifies the overall shift in scheduled times for the arrival flights. This calculation involved summing the time slots assigned to each flight, multiplied by 5 minutes (to convert from slots to actual minutes), and subtracting the original scheduled arrival time from the dataset. This approach ensures that the model considers how far each flight's new schedule deviates from its planned arrival time. A parallel calculation was performed for departure flights, ensuring alignment of new schedules with the original planned times.

Operational Constraints

To ensure the model remains feasible and meets operational requirements, a series of constraints were implemented to account for every flight and to avoid capacity breaches.

Flight Scheduling Constraints

Arrivals : For each arrival flight i , a constraint was introduced requiring that the sum of the binary decision variables w_arr[i, t] across all time slots t is at least 1. This guarantees each arrival flight is scheduled in at least one time slot, ensuring no flight is omitted. Departure: An analogous constraint was applied to departure flights j, ensuring each flight has a scheduled departure time.

Timing Adjustment Constraints

For each arrival flight, a constraint was defined to relate the difference between scheduled and rescheduled times to the displacement variables u_plus and u_minus. Specifically, the change in time equates to the difference between positive and negative displacements, accurately reflecting the rescheduling adjustments. A similar constraint was enforced for departure flights, maintaining consistency in scheduling adjustments.

Capacity/Runway Constraints

For each time slot t, a cap was set on the total number of flights that can be scheduled for arrival or departure. This limit, set to a maximum of 10 for both arrivals and departures per time slot, aligns the model with the operational capabilities of Atlanta Airport (ATL), ensuring manageable traffic flow and preventing capacity overload.

Displacement Constraints

To limit the extent of rescheduling adjustments, specific constraints were added for the positive and negative displacements of both arrival and departure flights.
Positive Displacement: The model constrains the positive displacement variable u_plus for each arrival flight iii to a maximum of 60 minutes, ensuring flights are not rescheduled too far forward in the day.
Negative Displacement: Similarly, the negative displacement u_minus for each arrival flight i is capped at 15 minutes, limiting how early an arrival flight can be shifted.

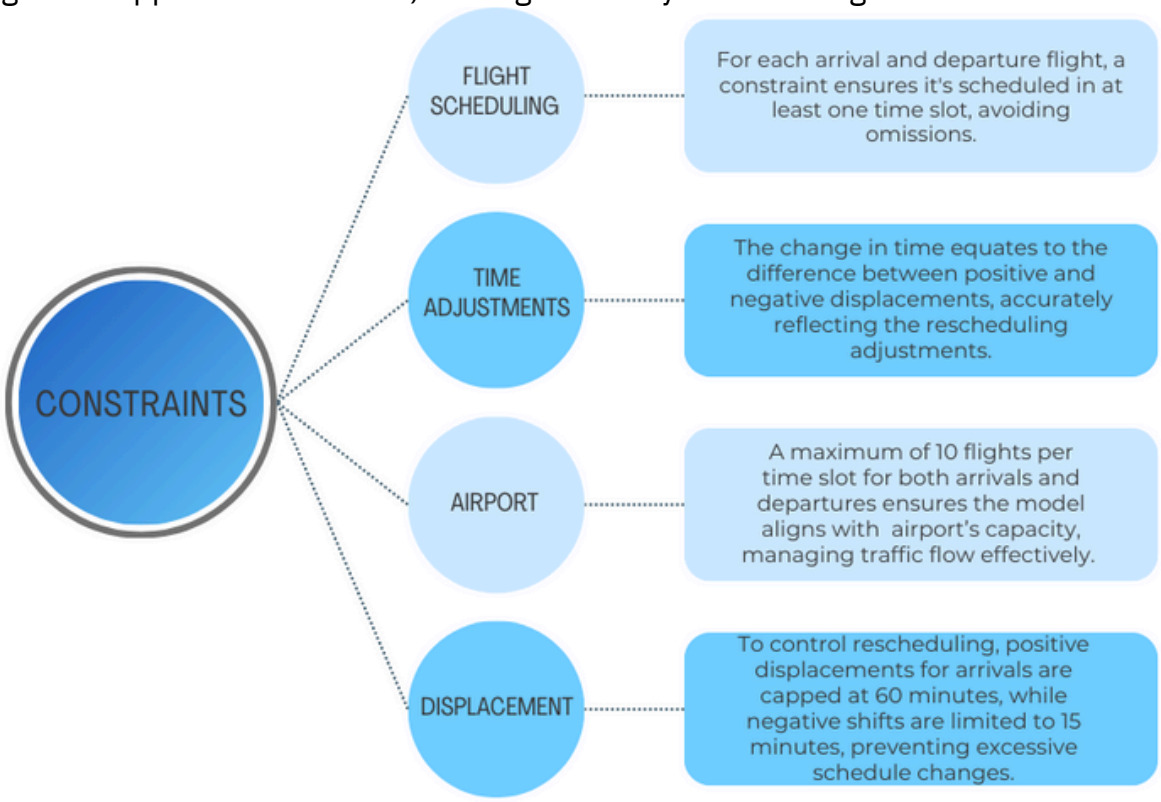

Fig-13: Constraints

Evaluation of Delay Reduction

After optimizing the flight scheduling model, we first calculated the total initial delays by summing the ARR_DELAY for arrivals and DEP_DELAY for departures at Atlanta Airport (ATL) to establish a baseline.

Following optimization, we assessed the new schedule to compute optimized delays. For each arrival flight, we identified the first scheduled time slot and calculated the delay by comparing the scheduled time (5 minutes multiplied by the slot index) to the original CRS_ARR_TIME, ensuring only positive delays were counted. A similar process was applied for departure flights.

We then summed the initial and optimized delays to find total values and calculated the reduction in delay by subtracting the total optimized delay from the total initial delay. The results highlighted the effectiveness of the scheduling adjustments on overall flight delays.

Result

Initial Total Delay: 553,817 minutes
Optimized Total Delay: 257199 minutes
Reduction in Delay: 296618 minutes
Percentage reduction in delay: 53.56%

The optimization model effectively reduced flight delays, showcasing its potential to enhance scheduling efficiency at Atlanta Airport. This substantial decrease in total delay highlights the model's capability to minimize disruptions and improve operational performance.

CONCLUSION

In this analysis, we developed a robust framework for predicting flight delays and optimizing flight schedules at Atlanta Airport. Starting with a dataset of approximately 1.3 million records derived from web scraping, we conducted extensive data preprocessing and exploratory analysis to identify key features and manage outliers. Our modeling efforts included multiple algorithms, with CatBoost emerging as the best performer, achieving an R^2 score of 0.918 after hyperparameter tuning. This highlighted the significance of fine-tuning parameters to enhance predictive accuracy.

Additionally, we utilized Gurobi's optimization capabilities to address the complexities of flight rescheduling. By formulating a model that minimized delays while adhering to operational constraints, we achieved a substantial 53.56% reduction in total delays. This dual approach of leveraging predictive analytics and optimization showcases the potential for improving efficiency and minimizing disruptions in airline operations, ultimately enhancing the overall performance of flight management systems.

ETHICS AND REPRODUCIBILITY

We provide the codebase for obtaining all the results along with the data curation pipeline. Further, we also validate that the data used does not contain any confidential information and is publicly available over the internet.