INDEX

Introduction

Airlines today face a range of operational challenges, from unpredictable delays to constrained resources, all of which impact profitability and customer satisfaction. Efficient flight scheduling, resource allocation, and quick turnaround are essential yet difficult to achieve consistently. Delays, in particular, lead to financial losses and disrupted schedules, making proactive solutions vital.

Data-driven optimization offers a promising solution by leveraging machine learning and analytics to anticipate issues and enhance decision-making. Predictive models can forecast flight delays using historical data on schedules, delay reasons, and crew availability, allowing airlines to adjust resources proactively. Furthermore, with advancements in technology, airlines can implement optimization strategies to reschedule flights effectively, reducing delays while meeting operational constraints.

However, the dynamic nature of aviation also suffers from the uncertainty of flight delays due to various factors, such as adverse weather, low visibility, and structural issues. These challenges, exacerbated by air traffic congestion, impact both operational efficiency and passenger satisfaction, making delay management a priority.

Problem Overview

In the aviation industry, flight delays are a persistent issue that affects operational efficiency, increases costs, and negatively impacts customer satisfaction. Delays can stem from a range of factors, including:

- **Weather Conditions**: Bad weather, such as storms, low visibility, and high wind speeds, disrupts scheduled operations.
- **Operational Constraints**: Limited resources, such as crew availability and airport infrastructure, create bottlenecks.
- **Air Traffic Congestion**: High demand and limited airspace contribute to delays, especially during peak travel times.
- **Unexpected Issues**: Structural defects or maintenance problems can further complicate scheduling.

These challenges lead to congestion and extended delays, which hinder the growth and efficiency of the aviation industry. Flight delays remain a major pain point for passengers, impacting satisfaction with airline services. Therefore, predictive models for delay management and optimized rescheduling solutions are critical to improving operational efficiency and maintaining competitive advantage.

Approach

To predict flight delays, we followed a structured approach encompassing data collection, preprocessing, feature engineering, model selection, and evaluation:

1. Data Collection

The first stage involved gathering data from various sources and selecting the most comprehensive dataset covering key features essential for accurate flight delay predictions. This dataset included factors like weather data, airline and airport performance, and operational metrics.

2. Exploratory Data Analysis (EDA) & Preprocessing

Once the data was collected, we conducted EDA to identify important features and trends. We processed categorical data by encoding necessary variables, handled missing values by dropping rows with null values, and created a correlation matrix to examine feature interdependencies. Based on the correlation heatmap and other visualizations, we removed columns with redundant time-related information, as well as the status column. The status column initially indicated whether flights were active or canceled (1 for active, 0 for canceled). We focused on active flights, as there wasn't sufficient data for canceled flights, and subsequently removed the status column.Given the large number of features relative to the sample size, we experimented with Principal Component Analysis (PCA) to reduce dimensionality, but it resulted in a significant drop in model accuracy. Therefore, we opted to use the full feature set for model training.

3. Model Selection

To identify the best model for predicting delays, we tested a variety of machine learning algorithms, including:

Classical Models: We implemented several foundational models, such as Linear Regression, LassoCV, Ridge, BayesianRidge, and LassoLarsCV. These models rely on linear relationships between input features and the target variable, which makes them suitable for datasets with strong linear dependencies. However, due to the complexity of relationships within our dataset, these models underperformed in comparison to other approaches.

Neural Networks: We explored neural networks using the MLPRegressor (Multilayer Perceptron Regressor), which is particularly effective for complex, nonlinear relationships in large datasets. While this model outperformed classical models, its performance was constrained by the relatively small size of our dataset, which affected the model's learning capacity. Moreover, neural networks require higher computational power and time, as reflected in the processing time for MLPRegressor.

Bagging Regressor: BaggingRegressor is an ensemble machine learning method that improves airline delay predictions by combining multiple regression models, typically decision trees. It trains each model on different random subsets of flight data, capturing diverse patterns and reducing the impact of outliers. By averaging their predictions, BaggingRegressor decreases variance and prevents overfitting, resulting in more accurate and stable delay forecasts. This model was found to be the best in our case.

4. Model Evaluation

Based on model scores, **BaggingRegressor** and **Random Forest Regresso**r emerged as the top-performing models for delay prediction. These models effectively captured the complexities within the data and delivered strong predictive accuracy.

5. Delay Analysis and Rescheduling Strategy

After identifying the best-performing models, we focused on analyzing significant delays. We created a separate data frame for flights delayed by more than 15 minutes using the arrival delay and status columns. Again, canceled flights were excluded due to insufficient data. The objective was to isolate patterns and causes of significant delays, enabling proactive scheduling adjustments. This rescheduling model serves as a foundation for minimizing future delays by dynamically addressing operational constraints based on real-time and historical delay data.

We propose a systematic rescheduling strategy that targets flights with arrival delays exceeding 15 minutes—the average delay. This approach begins by filtering and sorting these delayed flights, then classifying delays to focus on non-weather-related issues because weather related issues are hard to predict and will require advanced forecasts to generate viable results. By utilizing a Random Forest model with features such as flight schedules, historical delays, crew availability, maintenance records, and airport traffic, we predict delay probabilities and magnitudes. For each delayed flight, the algorithm iteratively adjusts the departure time within a 3-hour window by calculating optimal time gaps and reassessing delays until improvements stabilize or a 200-iteration limit is reached. Finally, the rescheduling process is applied to all relevant flights in prioritized order, effectively minimizing overall delays through predictive modeling and iterative optimization.

This structured approach lays the groundwork for a robust delay prediction and rescheduling model to improve operational efficiency and enhance customer satisfaction in the aviation industry.

Data Collection:
We tried to collect the data from various sources like the Bureau of Transportation Statistics, The Directorate General of Civil Aviation, Mendeley, Research papers etc.
We checked various datasets from different sources and finally we selected the dataset which covers the most important features like Weather Data, Delay Data, Airline-Airport Performance and Operational Performances. The data we used captures a range of key features relevant to analyzing and predicting flight delays. These are:

Date and Timing Data:

- **Used Date**: The specific date on which the flight operated, facilitating day-specific analyses of delay trends.
- **Scheduled Departure and Scheduled Arrival**: The scheduled times for departure and arrival, helping to evaluate any deviations from planned schedules.
- **SDEP and DEP**: Integer representations of scheduled and actual departure times, useful for precise time-based computations.
- **SARR and ARR**: Integer representations of scheduled and actual arrival times, aiding in assessing discrepancies between scheduled and actual operations.
- **Departure Delay and Arrival Delay**: Delay durations in minutes for departures and arrivals, providing insights into delay propagation and operational impacts.
- **Status**: A numerical indicator of the flight's status, potentially representing on-time, delayed, or canceled flights, enabling categorical analysis of performance.

Flight Route Data:

- **From and To**: Codes for origin and destination airports, allowing for airport-specific and route-based delay analysis.
- **Distance**: The total distance of the flight route in kilometers, which can influence delay likelihood due to extended flight durations or longer operational requirements.

Airline Data:

- **Airline**: Identifies the carrier operating the flight, supporting an airline-level view of delay patterns.
- **Airline Rating**: A numerical evaluation of the airline's performance, likely reflecting customer satisfaction or operational efficiency, which may correlate with delay patterns.

- **Market Share**: The airline's market share, providing insights into competitive dynamics on specific routes, which could impact scheduling flexibility.
- **Passenger Load Factor**: The percentage occupancy rate of the flight, potentially impacting boarding and deplaning times and therefore associated with delays.

Airport Data:

- **Airport Rating**: A performance rating for the airport, possibly reflecting efficiency, infrastructure quality, or operational capacity, which could correlate with delay frequencies.

Operational Performance Metrics:

- **OTP Index**: The On-Time Performance Index, which quantifies the frequency of on-time departures and arrivals, a critical metric for assessing delay trends.

Weather Data:

- **weather__hourly__windspeedKmph**: The hourly wind speed in kilometers per hour, as strong winds can affect flight schedules.
- **weather__hourly__weatherDesc__value**: A textual description of the weather (e.g., clear, cloudy), providing qualitative insights into environmental conditions that may impact flight operations.
- **weather__hourly__precipMM**: Hourly precipitation in millimeters, with rain or snow potentially contributing to delays due to safety requirements.
- **weather__hourly__humidity**: Humidity levels, which can influence flight operations, especially in cases of extreme weather conditions.
- **weather__hourly__visibility**: Visibility readings, essential for determining operational viability under low-visibility conditions.
- **weather__hourly__pressure**: Atmospheric pressure readings, which can affect takeoff and landing, especially in abnormal conditions.
- **weather__hourly__cloudcover**: Cloud coverage percentage, influencing visibility and potentially affecting delay likelihood.

Category:

- **Category**: A classification field, potentially used for segmenting or categorizing data, which can support additional layers of analysis for delay prediction.

Data Preprocessing & Feature Engineering :

The data cleaning process began by handling missing values. There were a few missing values in the dataset and the missing rows were removed. Columns with date and time, such as **Used Date**,

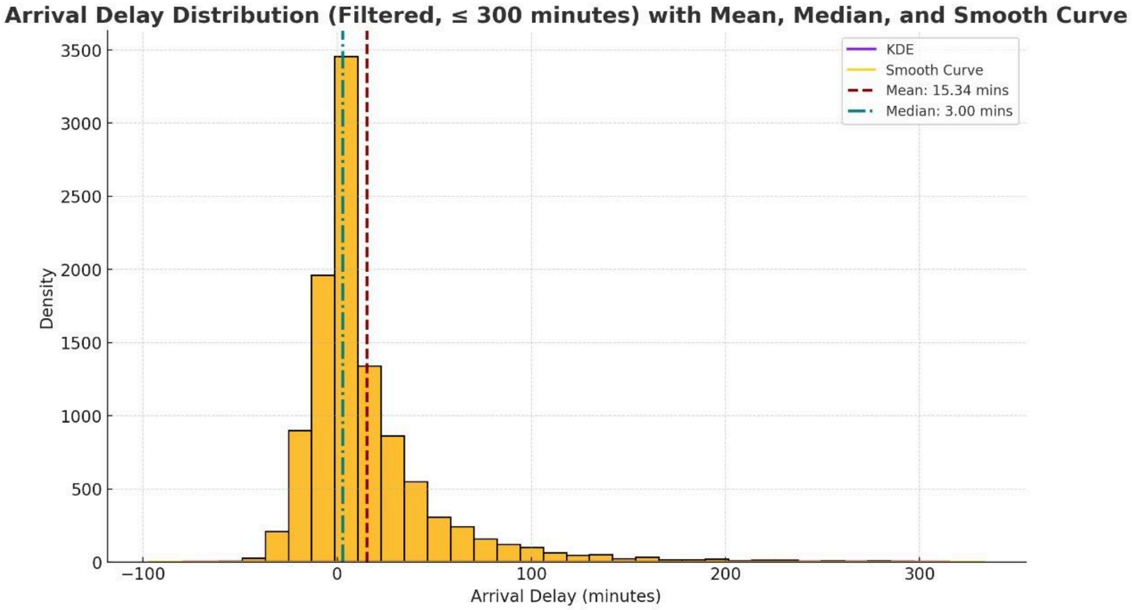
Scheduled Departure, and **Scheduled Arrival**, were standardized in Date-Time format to facilitate temporal analysis.

Descriptive Statistics

Summary statistics (mean, median, and standard deviation) were calculated for delay-related columns like **Departure Delay** and **Arrival Delay**, providing a foundational understanding of delay durations and identifying extreme values. Additionally, distributions of categorical data like **From**, **To**, and **Airline** were examined to understand the frequency and dominance of specific routes and carriers. One Hot Encoding was done for categorical features.

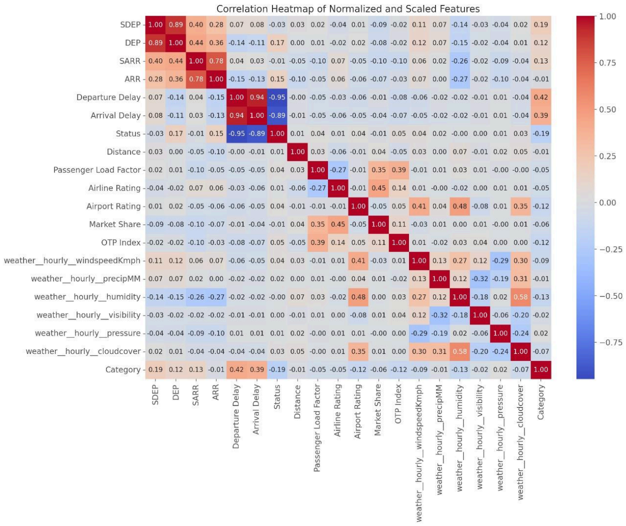
Feature Distribution Analysis

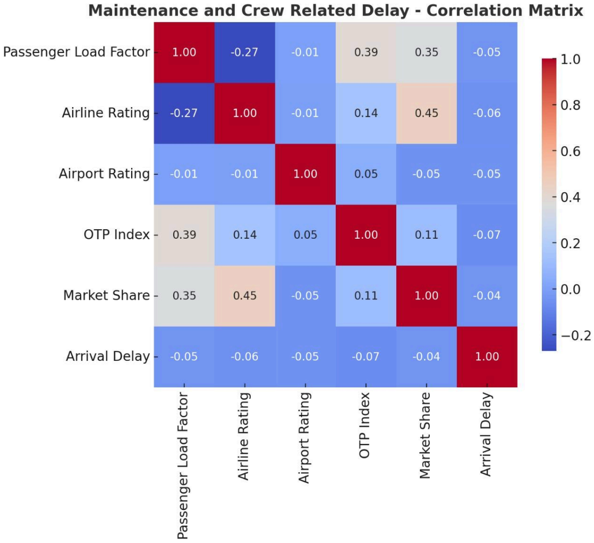
Understanding feature distributions is crucial for identifying patterns and biases. Graphs for continuous features, such as **Departure Delay**, **Arrival Delay**, **weather__hourly__windspeedKmph**, and **weather__hourly__humidity**, were plotted to show distributions, skewness, and outliers. The continuous features were also scaled appropriately.

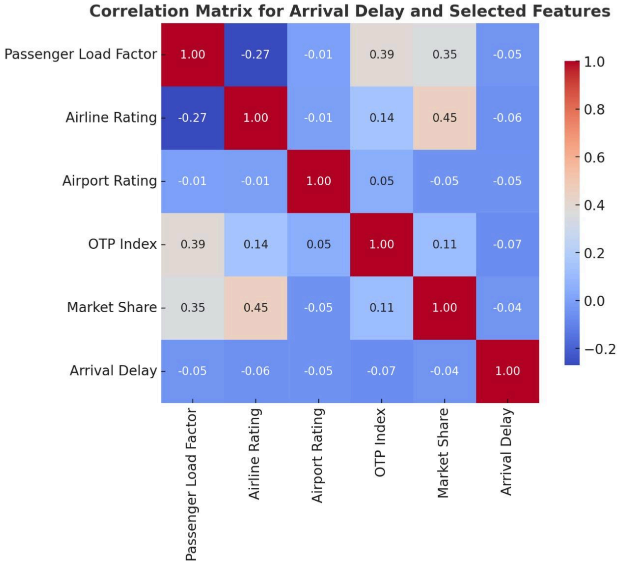


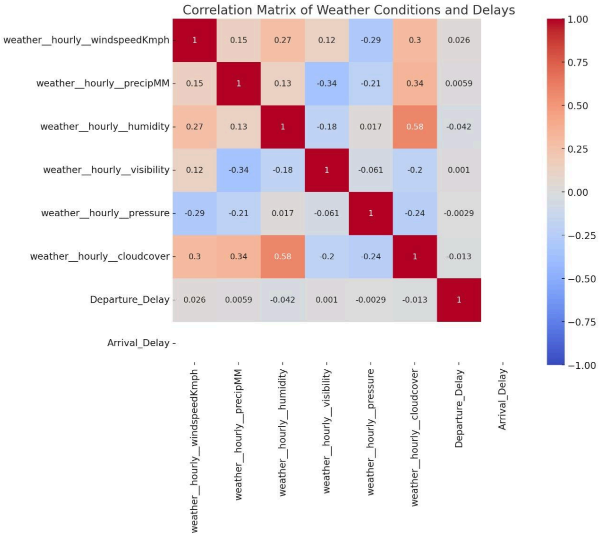
Correlation Analysis

To explore relationships among numerical features, correlations were calculated between variables such as Departure Delay, Passenger Load Factor, weather__hourly__windspeedKmph etc. A correlation heatmap visualized these relationships, highlighting features strongly related to delays. This analysis helped identify influential factors contributing to delays.







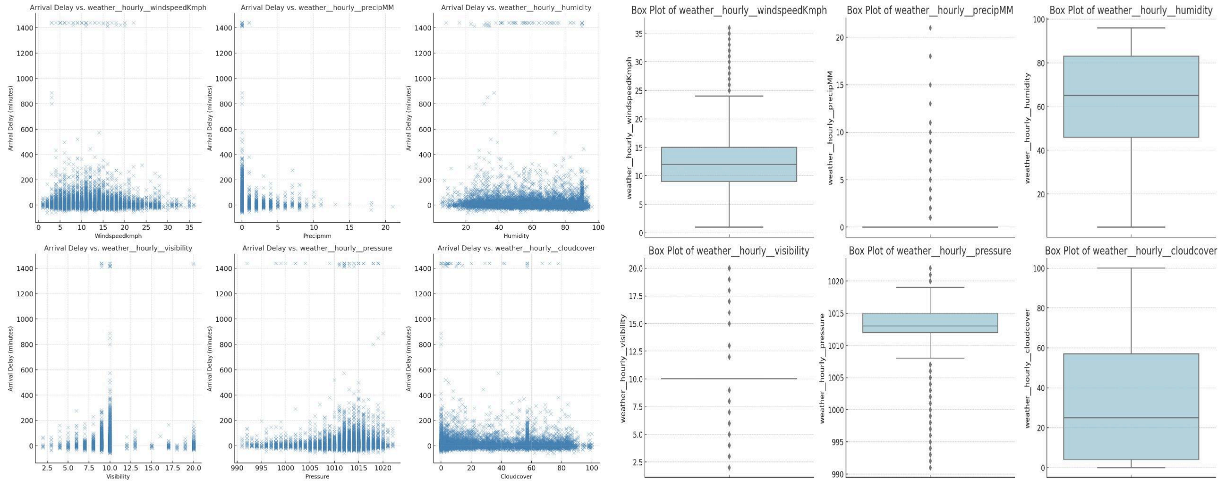


Time-Based Analysis

Temporal patterns in delays were assessed by examining delay data across intervals such as day of week, month, and Scheduled Departure time. By analyzing delays over these time dimensions, we identified peak delay periods on specific days or hours. These patterns were visualized by plotting average delays for each hour, day, and month, revealing seasonal and hourly trends.

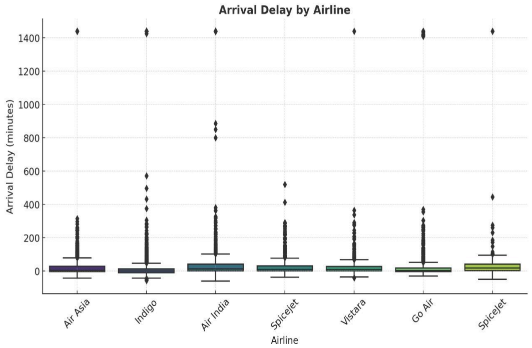
Weather Impact Assessment

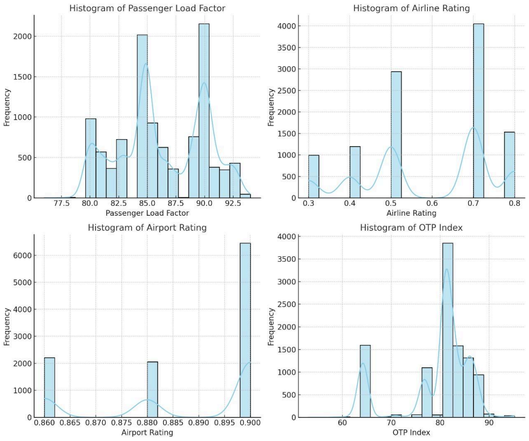
Weather-related features, such as weather__hourly__windspeedKmph, weather__hourly__precipMM, and weather__hourly__humidity, were examined to understand their impact on delays. Correlations between these features and delay variables (Departure Delay, Arrival Delay) were assessed. Boxplots compared delays under various weather conditions, such as high vs. low wind speeds or contrasting conditions like rain vs. clear skies, offering insights into weather's influence on delays. Mapping was done in the weather__hourly__weatherDesc__value column based on the severity of the weather to encode the categorical feature.



Route and Airline Analysis

Finally, delays across routes and airlines were analyzed by grouping data by From, To, and Airline. This helped identify specific routes and carriers with higher delay frequencies. Visualizations by airline and route pinpointed problematic routes or carriers, providing actionable insights for operational improvements in areas with frequent delays.





Models Implemented :

The following machine learning models were applied to predict and analyze flight delays, leveraging the diverse features in the dataset to address various aspects of delay prediction.

1. Linear Regression

- **Purpose**: Linear regression serves as a baseline model for predicting continuous outcomes like Departure Delay and Arrival Delay. It models a linear relationship between delay times and features such as weather conditions, time of day, and passenger load factor.
- **Advantages**: Interpretable and quick to train, making it useful for understanding basic trends in the data.
- **Limitations**: Struggles with complex, nonlinear relationships, limiting its ability to fully capture the range of delay influences.

2. Random Forest

- **Purpose**: Random Forest is widely used for delay prediction due to its robustness with large datasets and diverse features. By constructing multiple decision trees and averaging their predictions, it captures complex interactions between variables.
- **Advantages**: Handles categorical and numerical data, provides feature importance, and is generally robust to overfitting.
- **Limitations**: Computationally intensive, especially on large datasets, and may be unsuitable for real-time predictions.

3. Gradient Boosting Machines (GBM)

- **Purpose**: GBMs, including XGBoost and LightGBM, build models by combining multiple weak learners sequentially, each correcting errors from the previous one. This model effectively captures complex interactions between variables for predicting delays.
- **Advantages**: High accuracy, particularly with optimized hyperparameters like learning rate, depth, and number of trees.
- **Limitations**: Requires longer training times and careful tuning to avoid overfitting.

4. BaggingRegressor

- **Purpose**: Combines multiple instances of the same model (usually decision trees) trained on different random subsets of the data to reduce variance and improve stability in predictions.
- **Advantage**: Reduces overfitting by averaging the predictions from multiple models, making it robust on high-variance datasets.
- **Limitation**: Less effective on datasets where the underlying relationships are weak or noisy, and it doesn't improve much if the base model isn't strong.

5. XGBRegressor

- **Purpose**: An efficient implementation of gradient boosting designed to handle complex relationships by training sequentially on residuals (errors) of previous models.
- **Advantage**: Highly accurate and fast, with strong performance in handling both structured and unstructured data. Features regularization options to prevent overfitting.
- **Limitation**: Computationally intensive, requiring careful tuning of hyperparameters to avoid overfitting, especially on small datasets.

6. AdaBoostRegressor

- **Purpose**: Uses a sequence of weak learners (often decision stumps) that focus on correcting the errors of previous models, creating a strong predictor by emphasizing harder-to-predict cases.
- **Advantage**: Effective for improving accuracy with minimal adjustments, especially useful when simpler models underperform.
- **Limitation**: Sensitive to outliers and noise, as it places higher weight on misclassified points, potentially leading to overfitting.
-

7. Neural Networks (Deep Learning Models)

- **Purpose**: Neural networks, especially deep neural networks, can learn complex, hierarchical relationships within delay data, making them suitable for large datasets with numerous features, such as weather patterns and airport ratings.
- **Advantages**: Excellent for capturing intricate data patterns, offering high accuracy potential.
- **Limitations**: Resource-intensive, requiring large amounts of data and computational power, and generally less interpretable than tree-based models.

The different scores of all the implemented models are as follows:

Model	**Adjusted R-Squared**	**R-Squared**	**RMSE**
BaggingRegressor	**0.9661524556232209**	**0.9665789062523265**	**19.513566008334678**
RandomForestRegressor	**0.9626515074630269**	**0.9631220670983504**	**20.49791073006251**
ExtraTreesRegressor	**0.962263529914206**	**0.9627389777407653**	**20.60410219713155**
XGBRegressor	**0.9621232544091338**	**0.9626004695892334**	**20.64236767625923**
LGBMRegressor	**0.9454766265781511**	**0.9461635752866858**	**24.76648753155644**
HistGradientBoostingRegressor	**0.9454678460195383**	**0.9461549053557363**	**24.768481674868422**
DecisionTreeRegressor	**0.9398953313405086**	**0.9406525996810622**	**26.003221916590032**
GradientBoostingRegressor	**0.9395383590758823**	**0.9403001249671334**	**26.08032647021062**
KNeighborsRegressor	**0.936713108094016**	**0.937510469774586**	**26.682710386729635**
MLPRegressor	**0.9346856567785087**	**0.9355085626427085**	**27.106743155408015**
AdaBoostRegressor	**0.9319129515175149**	**0.9327707911390861**	**27.676126872107357**

Categorizing the Delay cause

We used the Bayesian inference to predict the likely cause of delays for each row in the test data, based on specific proxies associated with delay reasons like "Weather-related," "Operational," and "Route-related."

Using a Bayesian Network model trained on flight delay data, this analysis aims to predict the most probable cause of delay for each flight. The causes of delay are categorized into three major groups:

- **Weather-related**: Factors such as wind speed, precipitation, visibility, and humidity.
- **Operational:** Airline rating, OTP (On-Time Performance) Index, and airport rating.
- **Route-related:** Specific routes and distances, particularly major city routes (e.g., flights from BOM and DEL).

For each row in the test data, the model calculates a probability score for each delay reason based on the observed values in that row, then determines the most probable reason.

1. Model Training and Inference:

- The Bayesian Network is trained with `MaximumLikelihoodEstimator`.
- Inference is performed using the Variable Elimination algorithm to assess probabilities for each row.

2. Delay Reason Proxies:

- Weather, operational, and route factors are mapped to specific nodes in the Bayesian Network to assess the likelihood of each delay cause.

3. Probability Calculation :

- For each delay reason, a group probability is calculated by multiplying individual node probabilities from the proxy list.
- The delay reason with the highest group probability is chosen as the most likely cause for that row.

Key Insights:

- **Weather-related Delays :** Common for flights facing high wind speeds, precipitation, or low visibility.
- **Operational Delays :** Often linked to lower airline ratings, poor OTP scores, and lower airport ratings, indicating systemic or scheduling issues.
- **Route-related Delays :** Predominantly seen in specific routes with high demand or long distances, notably between major hubs.

This Bayesian inference approach offers a structured way to classify delays based on likely causes, which can guide operational improvements and preventive scheduling adjustments.

This analysis highlights the potential of data-driven approaches to improve airline operations by enabling proactive, data-informed scheduling and resource allocation adjustments, ultimately enhancing customer satisfaction and operational efficiency.

Rescheduling :

The first strategy that we used to approach the problem, was :

To effectively manage delayed flights, we can first predict potential delays, classifying them by underlying causes and saving this data in a separate CSV file for clarity. By identifying the primary reasons for delays—such as air traffic congestion, crew availability, and weather conditions—we can tailor our rescheduling strategy accordingly.

1. **Permanent Adjustments for Consistent Delay Causes:**

- For delays due to **air traffic congestion** or **crew unavailability**, we can analyze historical data on flights departing from the origin airport to identify peak congestion periods.
- Shifting flight schedules to off-peak hours, where traffic, crew availability, and maintenance resources are more consistent, can mitigate delays caused by these recurring issues. This proactive approach allows us to address two major factors contributing to consistent delays, reducing the need for continuous rescheduling.

2. **Dynamic Adjustments for Weather-Related Delays:**

- For **weather-related delays**, dynamic adjustments can be made 3-4 hours before departure based on the latest weather forecasts and industry norms.
- Rescheduling options should consider traffic conditions at both the departure and destination airports to prevent cascading delays. Flights can be adjusted either earlier or later, depending on operational needs and airport capacity at that time.

3. **Optimizing for Low Load Factor Flights:**

- For flights with **lower passenger loads**, flexible scheduling can be implemented to further reduce delays. By adjusting these flights to times with fewer expected disruptions, we can minimize overall delays without significantly impacting passenger experience.

Drawbacks of the Initial Approach: Shifting flights to off-peak hours isn't always feasible for high-demand routes or busy airports. Additionally, while effective for predictable delays, this method struggles with unexpected events like sudden weather changes or technical issues that require quick decisions.

Alternative Strategy: Systematic Rescheduling Using Predictive Modeling and Iterative Optimization

1. **Data Filtering and Preparation:**
    - **Threshold Selection:** Extract flights with arrival delays over 15 minutes, the average delay.

   - **Sorting:** Order these delayed flights in descending order of delay magnitude to prioritize rescheduling.
2. **Delay Classification:**
   - **Categories:** Classify delays into Weather-Related, Crew and Maintenance-Related, and Operational.
   - **Filtering:** Focus on non-weather-related delays for rescheduling, as weather delays need different handling due to their unpredictability.
3. **Predictive Modeling with Random Forest:**
   - **Model Utilization:** Use a Random Forest model to predict the likelihood and extent of delays based on historical and real-time data.
   - **Feature Selection:** Include features like flight schedules, historical delay patterns, crew availability, maintenance records, and airport traffic to improve prediction accuracy.
4. **Rescheduling Algorithm for Non-Weather Delays:**
   - **Iterative Process for Each Delayed Flight:** a. **Window Selection:**
     - Define a 3-hour window post the current scheduled departure.
     - Identify the next nearest flight within this window based on departure time. b. **Time Gap Calculation:**
     - Calculate Δx = (Next Flight Departure Time - Current Flight Departure Time) - 15 minutes buffer. c. **Departure Time Adjustment:**
     - Propose a new departure time by adding Δx minutes to the current departure.
     - Predict the new arrival delay using the Random Forest model.
     - **Decision Making:**
       - If the predicted delay decreases, accept the new departure time.
       - Otherwise, keep the original departure time. d. **Iteration and Convergence:**
     - Repeat adjustments until a constant delay is achieved or a 200-iteration limit is reached.
     - Finalize the rescheduled departure time based on the outcome.
5. **Application to All Relevant Flights:**
   - Apply the rescheduling algorithm to each delayed flight in the prioritized list to optimize overall delay reduction.

Conclusion:

After evaluating various machine learning models, ensemble-based methods, particularly BaggingRegressor model and Random Forest Regressor, consistently demonstrated high performance in predicting flight delays. Among these, the **BaggingRegressor model** stood out as the most accurate, with an **R-Squared value** of **0.9662**, showcasing its strong predictive ability for flight delay patterns based on the analyzed features. These models excelled due to their capability to handle complex interactions and non-linear relationships within the data. Although neural networks and other classical models provided comparable results, they demanded more computational resources and large datasets to work properly and generate better results, making ensemble models the optimal choice for practical, high-accuracy predictions in our case.

The proposed systematic rescheduling strategy effectively addresses the limitations of shifting flights to off-peak hours by focusing on flights with arrival delays exceeding the average of 15 minutes. By filtering and prioritizing these delays, classifying them to target non-weather-related issues, and leveraging a BaggingRegressor model for accurate delay predictions, the approach ensures data-driven decision-making. The iterative adjustment of departure times within a defined window allows for continuous optimization, while the application of this method to all relevant flights prioritizes overall delay reduction. This comprehensive strategy not only minimizes total delays but also enhances operational efficiency and reliability for high-demand routes and busy airports, ultimately leading to improved passenger satisfaction and better management of airport traffic.

ANNEXURE

https://github.com/maneshwarS/Flight-Delay-Prediction

https://drive.google.com/file/d/1jR_w0Z1g03nB0BlQWSt-nrZ9NpmRDNx0/view?usp=drive_link

https://drive.google.com/file/d/1mbGI2n5lkO5Qaza3cnQG547Xpfd7lEue/view?usp=drive_link

https://drive.google.com/file/d/1d3Y7YSxjKSH80zBlJAKrcDd8dSOEzYYm/view?usp=drive_link

https://drive.google.com/file/d/1eNOinjizgYVod1DHTMuANiaazUfrX9PB/view?usp=drive_link

https://drive.google.com/file/d/1hOiYSjfvZsXf4aVN0naD9e5FLClhCSYF/view?usp=drive_link

https://drive.google.com/file/d/19hQIhKe6V4jhVvR_agfn4NXqMpQgIbx7/view?usp=drive_link

https://docs.google.com/spreadsheets/d/1prmMosVfJm5XBGsAV_WOvLp9u3IE5AFanytvHNBUBq8/edit?usp=drive_link