Optimizing Air Travel: A Data-Driven Approach to Flight Delay Analysis and Prediction

Project Report

Author

Tanuj Kumawat

(Enrollment no. 23112106)

Date

16 June 2025

1. Problem Statement / Objective

This project aims to predict delays, prioritize controllable causes (like crew/aircraft issues), and provide actionable insights—helping airlines reduce delays and improve efficiency.

2. Dataset Description

- Rows: 179338 records.
- Columns: 21 original features including, Airport , carrier details,count and values(in minutes) of delay causes,time (year,month) of flights for trend analysis(seasonal) , and the target variable (`arr_delay`).

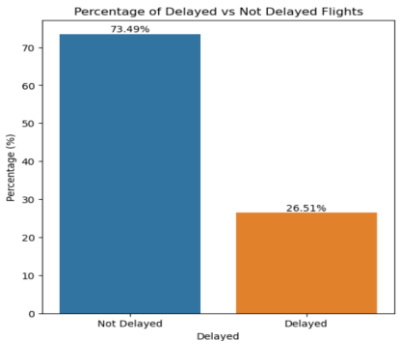
3. Data Cleaning & Preprocessing

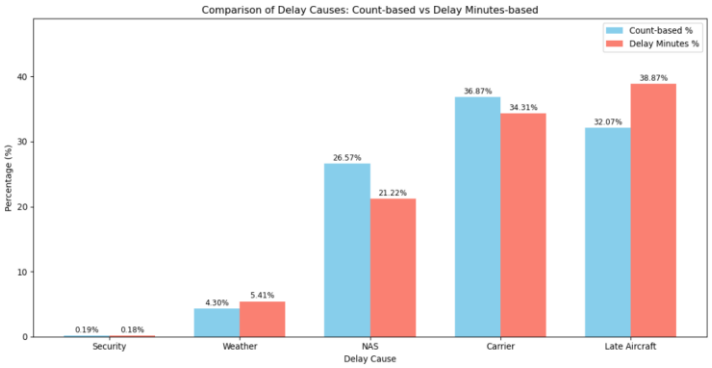
- There are 0 duplicated values and some null values in ['arr_flights', 'arr_del15', 'carrier_ct', 'weather_ct', 'nas_ct', 'security_ct', 'late_aircraft_ct', 'arr_cancelled', 'arr_diverted', 'arr_delay', 'carrier_delay', 'weather_delay', 'nas_delay', 'security_delay', 'late_aircraft_delay'] columns. Nulls in these columns are replaced with median values.
- Since the dataset given is for total flights arrived on a particular airport, specific carrier, in a month so to convert the dataset for a single flight by dividing all the numerical columns with the no of total flights ('arr_flight') for predicting, whether a flight is likely to be delayed (Yes/No) and estimating the expected delay duration (in minutes).
- We create a new column ('is_delayed') if ('arr_delay'>15) then flight is delayed(1) else not delayed (0).
- Drop ' airport_name' and 'carrier_name' as they have no impact on the model and analysis.

4. Exploratory Data Analysis (EDA):

- Univariate and bivariate analyses were performed.
- Bar and count plots for distributions.

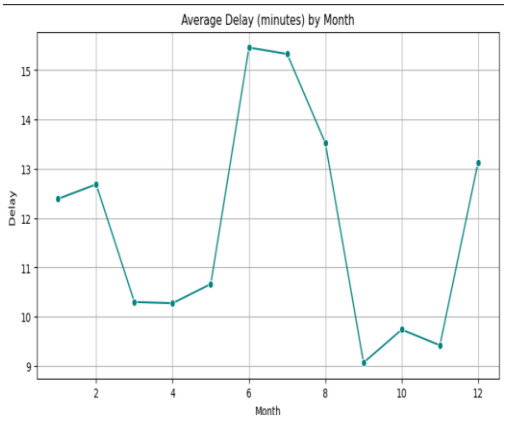
Univariate Analyses graphs and Result :

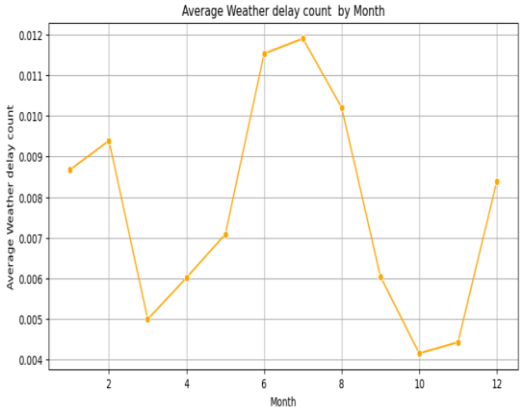


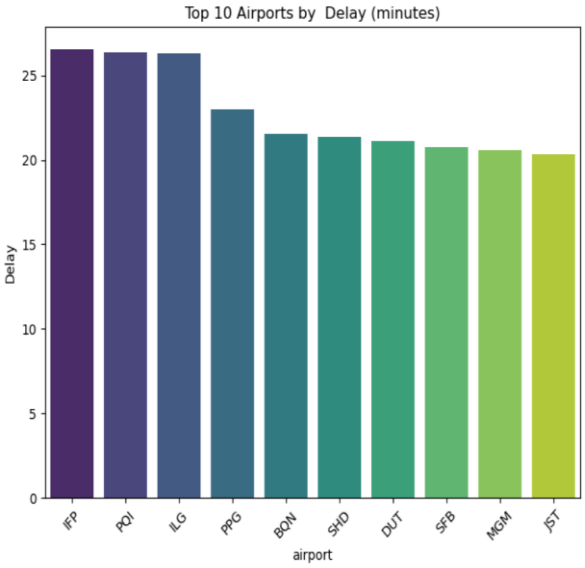


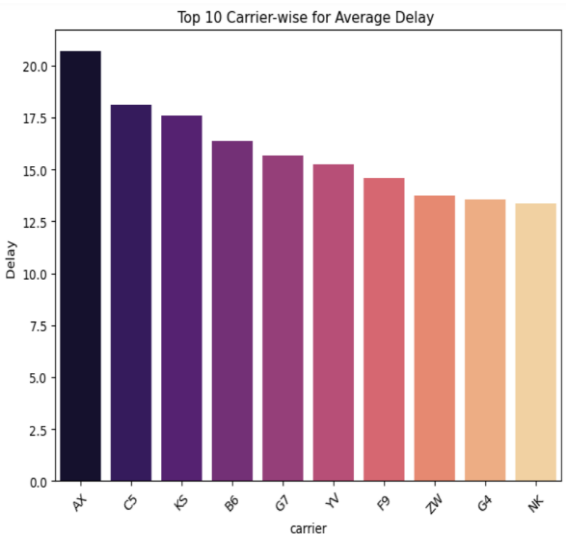
- Data is imbalanced with 26.5 % delayed flight.
- **Late Aircraft and Carrier Delays** are the **dominant contributors** both in count and delay minutes, suggesting a focus area for operational improvements, while security delay contribution is negligible . So we drop 'security_ct' and 'security_delay' variables.

Bivariate Analyses graphs and Results:









- "June" and "July" experience more weather-related delays due to thunderstorms, monsoons, and heatwaves. These weather disruptions significantly contribute to overall flight delays during this peak travel season."
- **Top Delayed Airports**: Airports like **IFP, PQI, ILG** experience the highest average delays, indicating potential operational or weather-related challenges at these locations.

- **Carrier Delays**: Airlines such as **AX, C5, KS** have higher average delays compared to others, suggesting either operational inefficiencies or coverage of delay-prone routes.

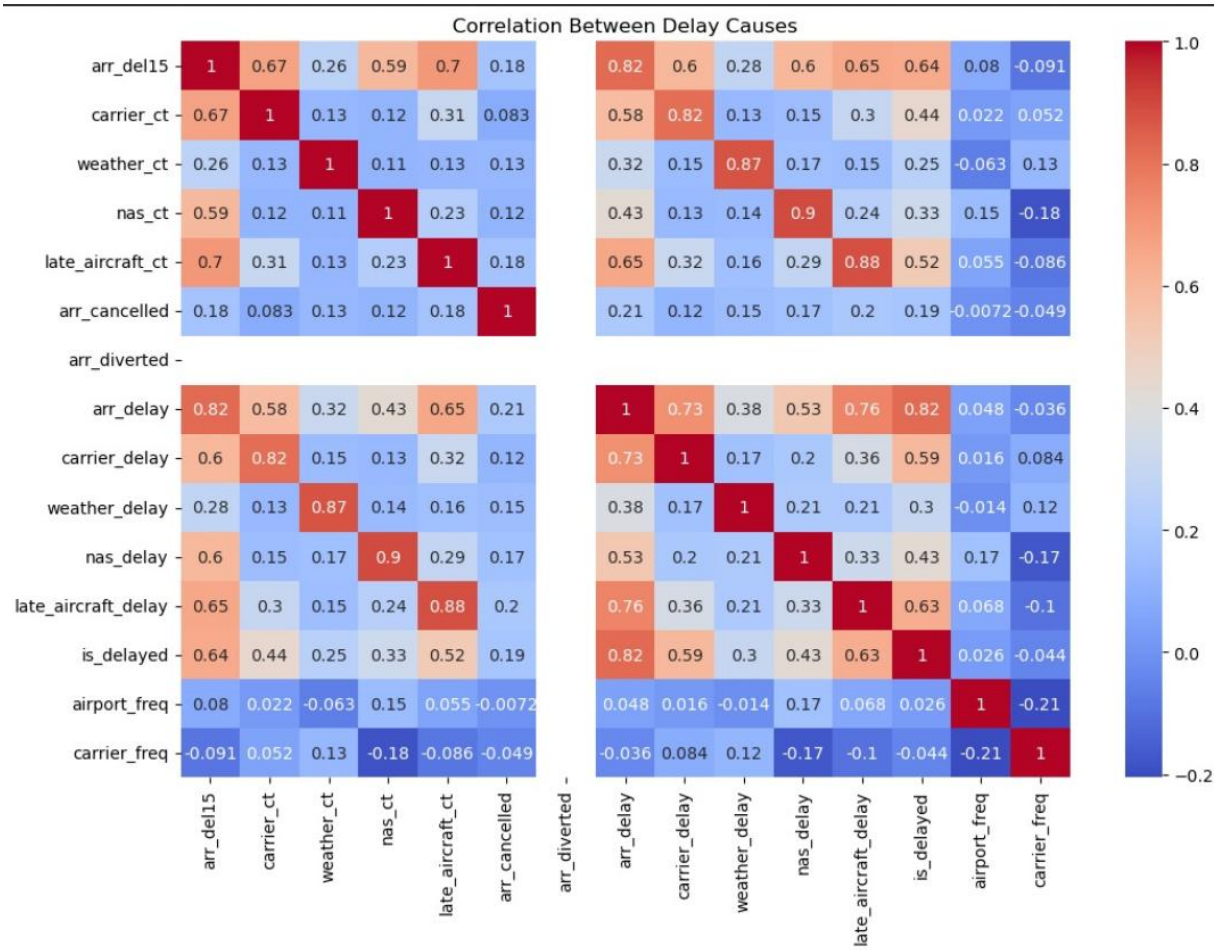
5.Feature Engineering :

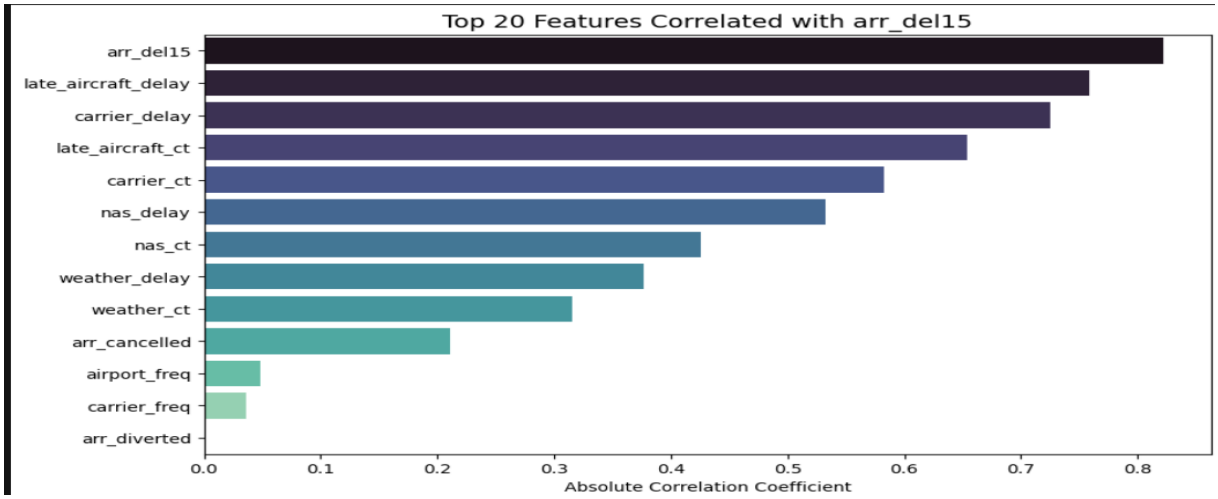
- Apply Frequency encoding for 'airport' and 'carrier' variables as it reduces the dimensionality of high-cardinality categorical variables like **'airport'** and **'carrier'**, making the data more model-friendly, improving computation speed, and preventing overfitting compared to one-hot encoding.Now new variables are 'airport_freq' and 'carrier_freq' and drop the old columns.
- For numerical features, the majority of distributions are right-skewed.. It demonstrates that these columns have many outliers.
- Most of the variables can get a normal distribution when outliers are handled by **Clipping Method**.

6. Feature Selection & Encoding:

Correlation Coefficient and Heatmap:

- The correlation coefficient is an important tool in data analysis and machine learning, as it can help to identify relationships between variables and can be used in feature selection techniques to remove the variables which has less correlation with the target variable.





- The arr_diverted variable has most of the values is 0 so no correlation with the target variable , so remove it .
- Apply ohe on 'month' and standard scaler method on all the numerical columns.

7. Handling Imbalance

'is_delayed' true values are very less (26.5%) compared to false values. This can be handled using SMOTE (Synthetic Minority Oversampling Technique). SMOTE was applied on training data (X_train, y_train). SMOTE made the distribution of 0 and 1 equal in the training split.

Before SMOTE:

0	**73.5%**
1	26.5%

After SMOTE:

0	**50%**
1	50%

8. Model Building for Classification:

Models used

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- LightGBM

Train-Test split of 80-20 was used. Then, Hyperparameter tuning was done for all the above mentioned models using GridSearchCV. The grid is in the project notebook.

- Scoring for tuning is done according to F2 score. That means we selected those hyperparameters which gave the most F2 Score.
- $F2\ Score = (1 + 2^2)\frac{Precision.Recall}{4Precision + Recall}$
- F2 Score gives more importance to Recall. Recall is sensitive to False negative predictions. F2 Score prioritizes recall to avoid missing actual delays, helping airlines proactively manage and minimize disruption and costs.

Following tabulates the best parameters:

Logistic Regression	{'C': 0.1, 'penalty': 'l2'}
Decision Tree	{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 2}
Random Forest	{'max_depth': 20, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 50}
XGBoost	{'learning_rate': 0.1, 'max_depth': 7, 'min_samples_leaf': 6, 'min_samples_split': 5, 'n_estimators': 50}
Light GBM	{'max_depth': 7, 'min_samples_leaf': 6, 'min_samples_split': 4, 'n_estimators': 50}

9. Classification Model Evaluation

All models were evaluated on the basis of ScikitLearn's Classification report, train and test accuracy,Recall,F1score, ROC-AUC score and F2 score. Results are tabulated below.

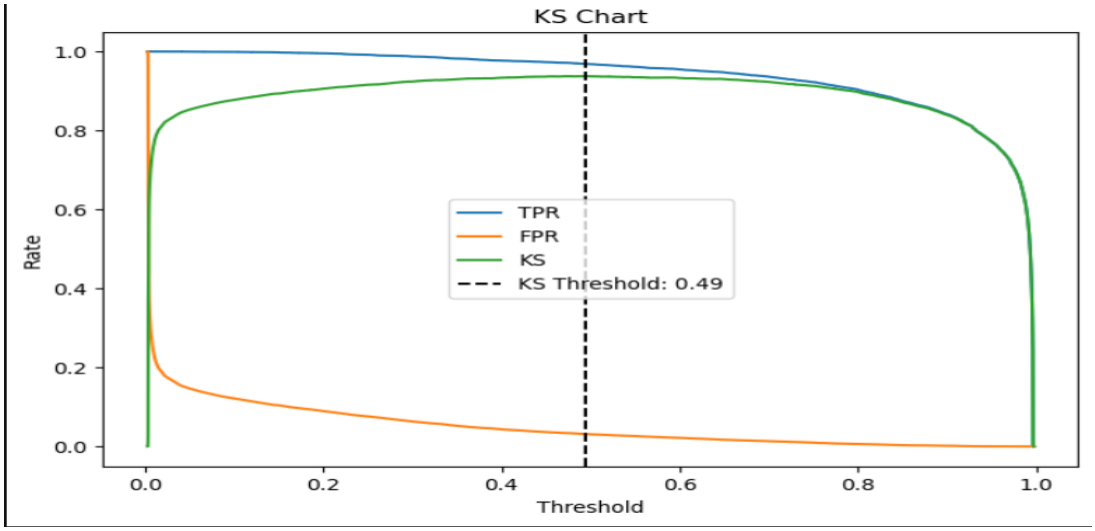
```
Final Classification Model Comparison:
                model  train_accuracy  test_accuracy  train_precision  \
3             XGBoost        0.976681       0.968914         0.972900
4            LightGBM        0.971393       0.965903         0.968761
2        RandomForest        0.992659       0.965317         0.990537
1        DecisionTree        0.969017       0.956702         0.964493
0  LogisticRegression        0.948848       0.951684         0.955291

   test_precision  train_recall  test_recall  train_f1   test_f1  train_f2  \
3        0.918286      0.980679     0.968980  0.976774  0.942952  0.979113
4        0.911838      0.974201     0.964669  0.971473  0.937510  0.973108
2        0.913548      0.994821     0.960042  0.992674  0.936218  0.993961
1        0.891238      0.973888     0.952997  0.969168  0.921083  0.971994
0        0.884277      0.941772     0.940904  0.948483  0.911712  0.944445

    test_f2  train_roc_auc  test_roc_auc
3  0.958398       0.998155      0.996770
4  0.953618       0.997194      0.996207
2  0.950368       0.999793      0.994970
1  0.939970       0.996725      0.988928
0  0.929006       0.988023      0.987552
```

- Xgboost selected as the final model based on superior Test F2-score (95.83%) and high Recall (96.89%), aligning with the business objective of minimizing False Negatives (missed detections).
- Thresholding based on KS-statistic optimized Recall-F2 trade-off.
- Threshold optimized for KS statistic, ensuring best separation between classes.

- Improved Recall & F2 Scores at the cost of slight drop in Precision — acceptable in high Recall-focused problem.



10. Model Building for Regression:

Models used

- Linear Regression
- Decision Tree
- Random Forest
- XGBoost
- LightGBM

Train-Test split of 80-20 was used. Then, Hyperparameter tuning was done for all the above mentioned models using GridSearchCV. The grid is shown below.

- Models are evaluated based on a **combination of low MAE, low RMSE, high R^2 Score, and high OAI**.
- The **best model** balances error reduction (MAE & RMSE) with **explainability (R^2)** and **operational relevance (OAI)** — ensuring predictions are not only accurate but also actionable for airline operations.

Following tabulates the best parameters:

Decision Tree	**{'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 2}**
Random Forest	{'max_depth': 15, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 50}
XGBoost	{'learning_rate': 0.1, 'max_depth': 7, 'min_samples_leaf': 6, 'min_samples_split': 5, 'n_estimators': 50}
Light GBM	{'max_depth': 7, 'min_samples_leaf': 6, 'min_samples_split': 4, 'n_estimators': 50}

11. Regression Model Evaluation:

Results are tabulated below:

```
Final Regression Model Comparison (Including OAI):
                   model  train_MAE  test_MAE  train_RMSE  test_RMSE
0      Linear Regression   1.167384  1.181978    2.313168   2.346006
1  DecisionTreeRegressor   1.222524  1.328538    2.002451   2.217291
2  RandomForestRegressor   0.521043  0.766665    1.272596   1.871314
3       XGBoostRegressor   0.763920  0.829891    1.705949   1.875591
4      LightGBMRegressor   0.872861  0.908502    1.825164   1.911241

   train_R2   test_R2       OAI
0  0.916137  0.913512  7.486956
1  0.937154  0.922742  7.442937
2  0.974617  0.944971  7.441417
3  0.954387  0.944719  7.443173
4  0.947789  0.942597  7.444177
```

LightGBMRegressor is selected as the best model because it offers a strong balance between prediction accuracy and operational relevance:

Low Test MAE (0.9085) and Test RMSE (1.9112) show that it generalizes well to unseen data, outperforming Decision Tree and Linear Regression.

It has a high test R^2 score (0.9426) — very close to the top-performing models, indicating reliable predictive power.

The OAI (7.389963) is comparable to the highest values, demonstrating its suitability for real-world airline operations.

12.Model Explanability using shap:

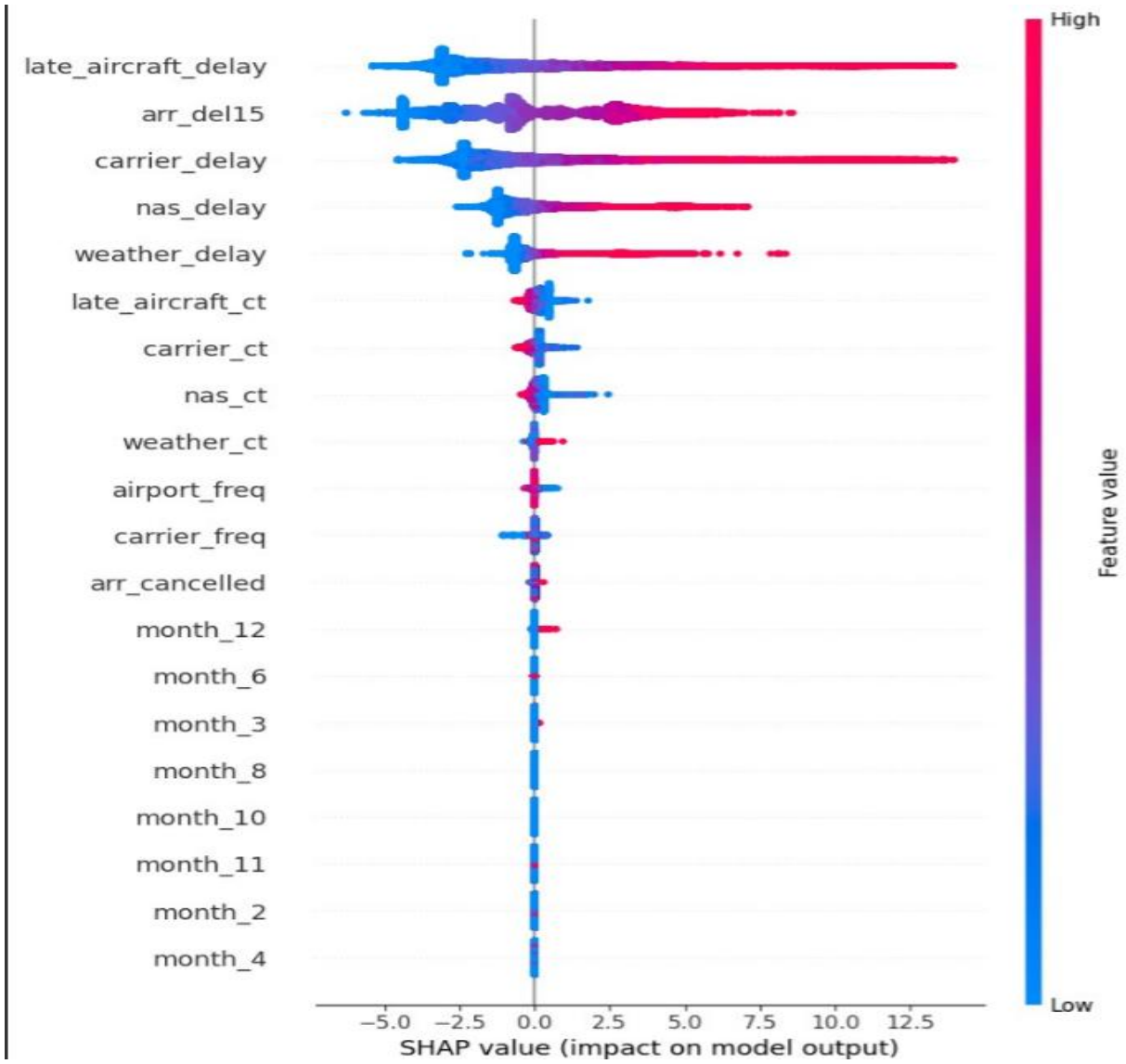
- The **most impactful features** (by SHAP value) are:
    - `late_aircraft_delay`
    - `arr_del15`
    - `carrier_delay`
    - `nas_delay`
    - `weather_delay`

- These features have the **largest SHAP values**, meaning they contribute most to the model's prediction.

Controllable vs. External Delays

- **Controllable delays** (e.g., `carrier_delay`, `nas_delay`, `late_aircraft_delay`) have a much **higher OAI-weighted SHAP score (2.1533)**.
- **External delays** (e.g., `weather_delay`) have a lower weighted score (**0.6037**)

Controllable delays contribute more significantly to overall flight delay prediction **and have a greater impact on airline operations**, as measured by the **Operational Actionability Index (OAI)**.



13. Actionable Recommendations & Consulting Insights:

1. **Schedule Optimization:**
   Adjust flight schedules, especially for carriers and airports prone to peak-hour congestion, to reduce systemic delays.
2. **Enhance Ground Operations:**
   Invest in faster aircraft turnaround processes and efficient baggage handling to minimize carrier-related delays.
3. **Weather-Responsive Planning:**
   Integrate real-time weather forecasting in operations to preempt weather-driven disruptions and reassign resources proactively.

4. **Proactive Passenger Communication:**
   Implement AI-driven notifications to inform passengers about expected delays, improving customer satisfaction and trust.
5. **Resource Reallocation:**
   Prioritize deployment of backup aircraft and crew at high-risk hubs based on predicted delay probabilities to reduce late aircraft impacts.
6. **Collaborative Air Traffic Management:**
   Work with air traffic control for rerouting options during NAS (National Airspace System) constraints to avoid cascading delays.
7. **Performance Monitoring Dashboards:**
   Develop live dashboards showing OAI-driven delay factors, enabling operational teams to intervene in controllable delay causes.