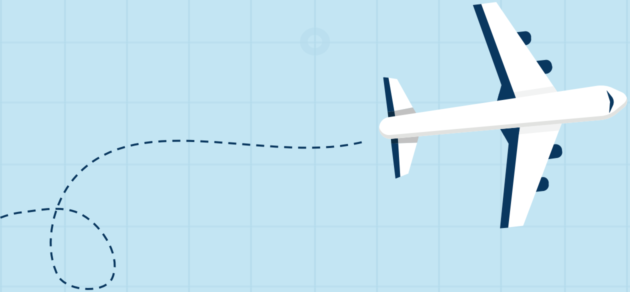


Flight Delay Prediction &Insights Open Projects 2025 Analytics

Bulla Reena Jasper
22118018

Dataset Overview

Dataset: `Airline_Delay_Cause.csv`

Key Features:
Flight Info,Delay Categories,Total delay and arrival delay columns

Target Variables: Classification: Will the flight be delayed? (Yes/No)
Regression: Delay duration (in minutes)

Preprocessing Steps:

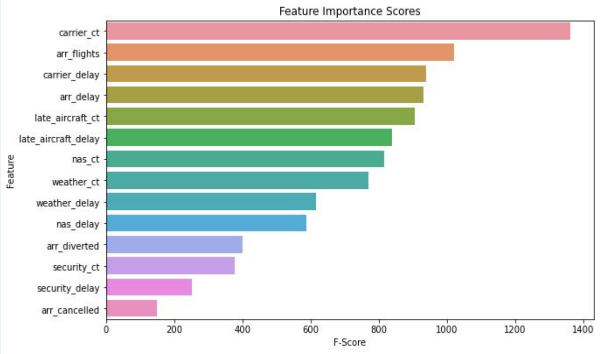
-Cleaned missing/null values
-Feature scaling using StandardScaler
-Label encoding for categorical variables
-Separation of **controllable vs uncontrollable** delays (for OAI)

EDA Insights (Exploratory Data Analysis)

- **What I Explored:** Class Distribution of delayed vs. non-delayed flights (is_delayed), Feature Importance for classification, Feature Importance for classification, Arrival Delay Duration distribution, Proportion of 15+ minute delays, Correlation Matrix for numeric features

- **Proportion of 15+ minute delays**

- **Correlation Matrix** for numeric features



- **Delay Imbalance**

-95% of flights are delayed by 15+ minutes

-Only ~5% are on-time or under the threshold

- **Highly Influential Features**

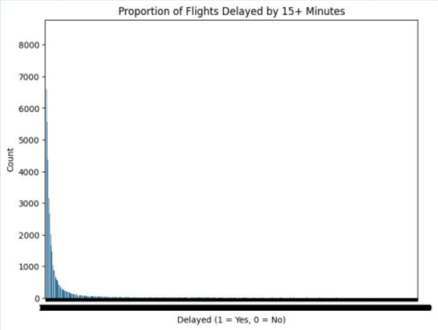
carrier_ct, arr_flights, carrier_delay, and arr_delay scored highest in importance

Most influential factors are controllable

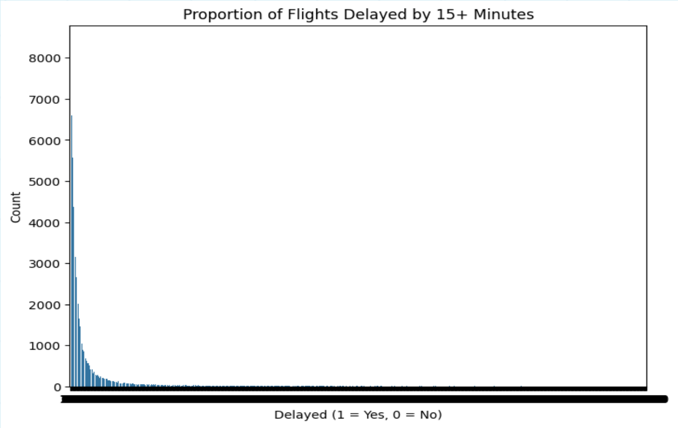
- **Delay Duration is Skewed**

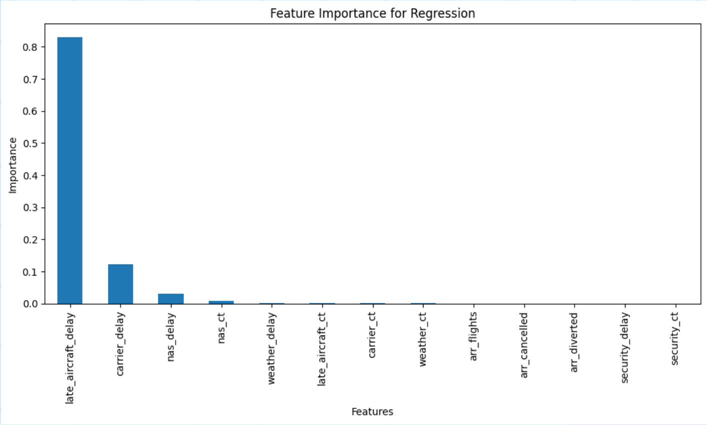
-A few extreme delay values inflate the distribution

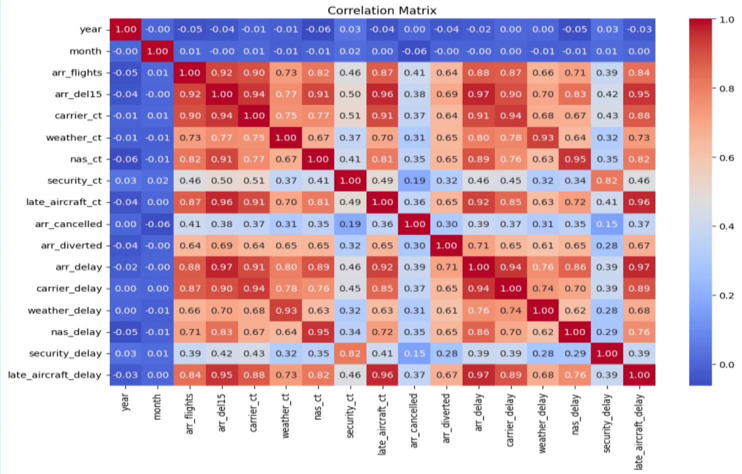
-Majority of delays are within 0–100 minutes



Some more insights from the Analysis







Model Performance (Both)

Classification Models (Predicting Delay: Yes/No)
Models Trained:
Logistic Regression,Random Forest Classifier, Support Vector Machine (SVM)

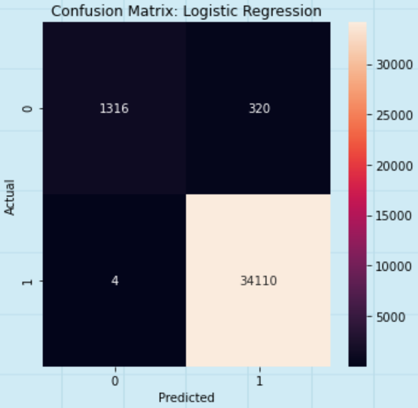
Model	CV Accuracy Mean	Std Dev
Logistic Regression	**0.9885**	±0.0012
Random Forest	**1.0000**	±0.0000
SVM	**0.9529**	±0.0012

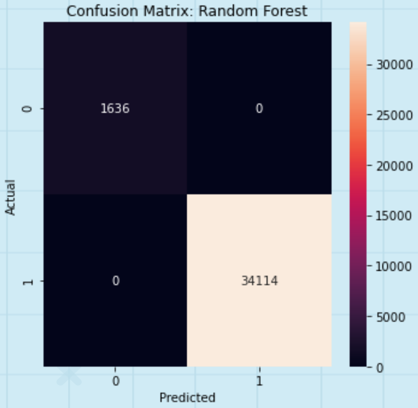
Regression Model (in min)

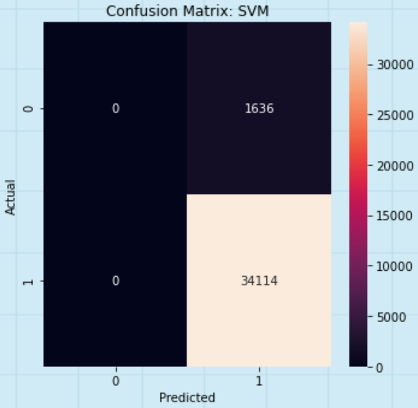
Model Used: Random Forest Regressor

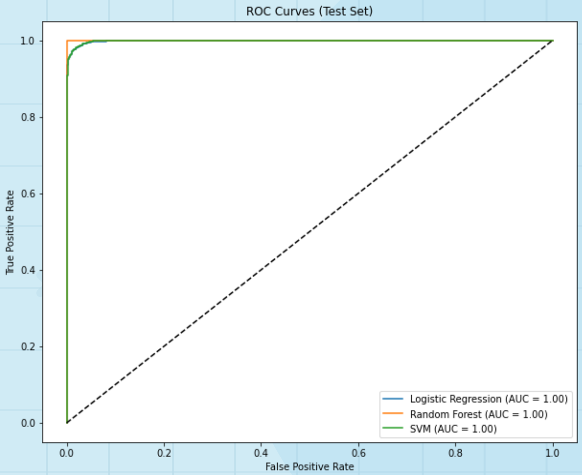
Metric	Value
RMSE	665.10 min
MAE	86.02 min
R^2 Score	1.00
Mean CV RMSE	795.09 min

Classification Model Evaluation – Metrics & Visuals







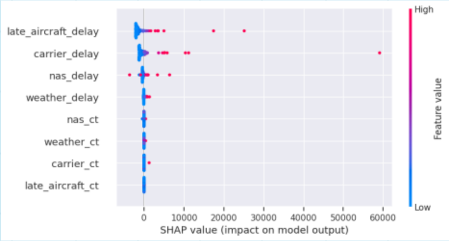


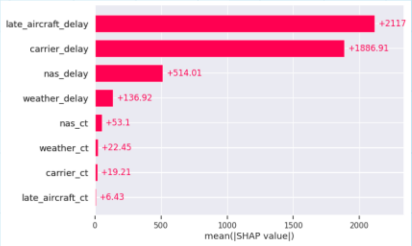
Model	**AUC Score**
Random Forest	**1.000**
Logistic Regression	**0.9987**
SVM	**0.9987**

SHAP Insights – Classification & Regression Models

Classification (Will the flight be delayed?)

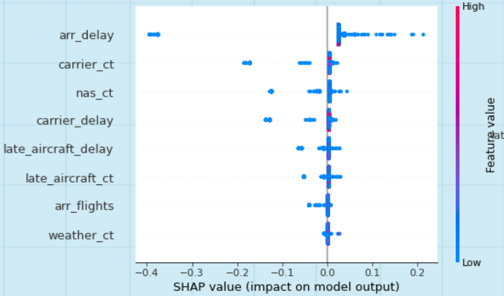
Top feature: arr_delay
High impact: carrier_ct, carrier_delay, nas_ct
SHAP shows controllable delays strongly influence "Delayed" prediction

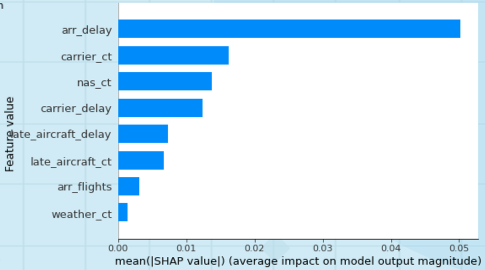




Regression (How long will the delay be?)

Top features: late_aircraft_delay, carrier_delay
Impact: Higher delay values consistently increase predicted delay duration
SHAP confirms operational delays (carrier, aircraft) are key drivers of delay minutes





Operational Focus – OAI & Recommendations

OAI was used to:

Prioritize controllable delays (e.g., carrier_delay, late_aircraft_delay)
Weight features in training and SHAP to **highlight actionable drivers**
Guide airlines to focus on delays **they can fix**, not weather or NAS-related issues
*About this more clearly it is mentioned in notebook

Recommendations:

Use Random Forest for deployment due to its robust performance.
Focus on reducing carrier_ct and late_aircraft_ct delays.
Reduce late aircraft delays with better turnaround scheduling
Optimize carrier operations (crew, gate, dispatch)
Avoid peak hours by retiming high-risk flights
Focus on high-delay airports for resource planning
Use model predictions to trigger early operational action

THANK YOU