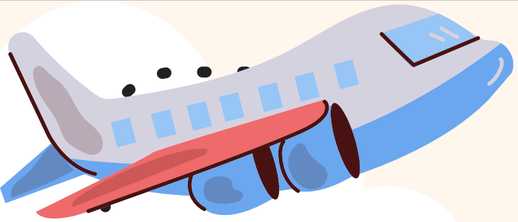
OPTIMIZING AIR TRAVEL: A DATA-DRIVEN APPROACH TO FLIGHT DELAY ANALYSIS AND PREDICTION

Presented by: Vaisant K , 22117147

PROJECT OVERVIEW & PROBLEM STATEMENT



AIR TRAVEL IS A CRITICAL PILLAR OF GLOBAL CONNECTIVITY, YET FREQUENT FLIGHT DELAYS DISRUPT SCHEDULES, REDUCE CUSTOMER SATISFACTION, AND INCREASE OPERATIONAL COSTS. AIRLINES STRUGGLE TO PROACTIVELY MANAGE DELAYS, ESPECIALLY WHEN THE CAUSES ARE DIVERSE AND OFTEN INTERDEPENDENT.

PROBLEM STATEMENT

"DEVELOP A PREDICTIVE SYSTEM THAT NOT ONLY DETERMINES WHETHER A FLIGHT WILL BE DELAYED BUT ALSO ESTIMATES THE DURATION OF THE DELAY. THE MODEL SHOULD GO BEYOND GENERIC FORECASTING BY INCORPORATING FACTORS THAT AIRLINES CAN INFLUENCE, SUCH AS CARRIER-RELATED AND LATE AIRCRAFT DELAYS. BY FOCUSING ON CONTROLLABLE CAUSES, THE SYSTEM AIMS TO SUPPORT OPERATIONAL DECISION-MAKING AND HELP AIRLINES PROACTIVELY REDUCE AVOIDABLE DISRUPTIONS."

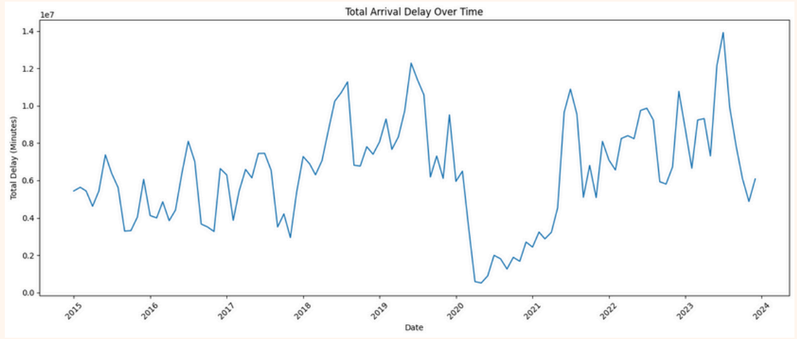
PROJECT OBJECTIVES

- UNCOVER DELAY PATTERNS
- PREDICT DELAYS
- PRIORITIZE CONTROLLABLE DELAYS
- RECOMMEND ACTIONS

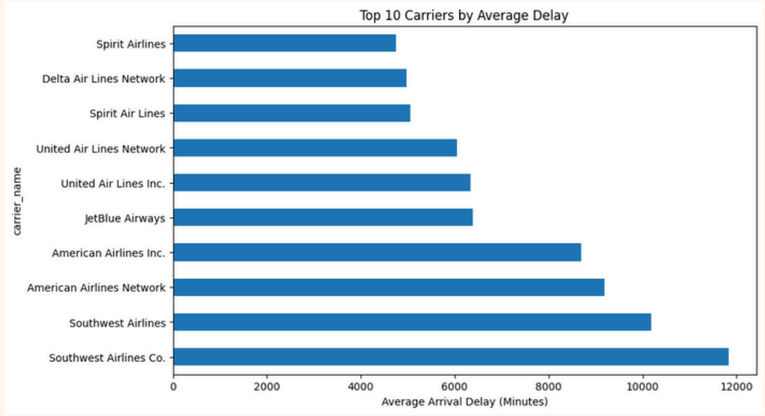
THIS PROJECT EMPOWERS AIRLINES TO PROACTIVELY MANAGE RESOURCES, MINIMIZE DELAY-RELATED COSTS, AND ENHANCE PASSENGER EXPERIENCE BY FOCUSING ON THE DELAYS THEY CAN ACTUALLY CONTROL.

EXPLORATORY DATA ANALYSIS (EDA)

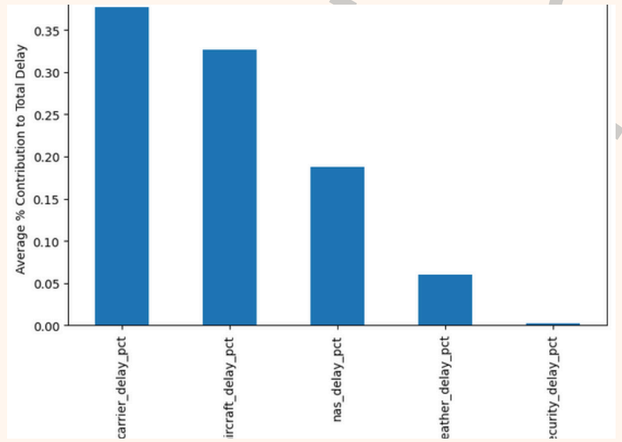
WE PERFORMED EDA TO UNDERSTAND DELAY TRENDS, IDENTIFY INFLUENTIAL VARIABLES, AND UNCOVER PATTERNS THAT DRIVE PREDICTIVE MODELING AND OPERATIONAL RECOMMENDATIONS.



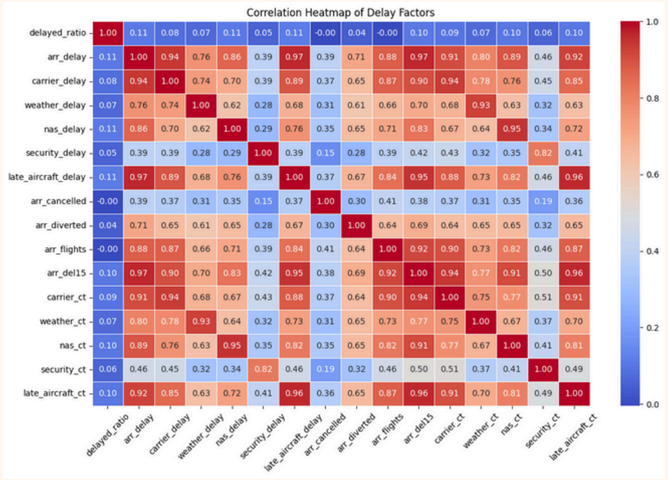
TOTAL ARRIVAL DELAYS SHOW SEASONAL PEAKS, A RISING TREND UNTIL 2019, A SHARP DROP DURING THE 2020 PANDEMIC, AND INCREASED VOLATILITY WITH NEW HIGHS IN 2022-2023.



SOUTHWEST AIRLINES HAS THE HIGHEST AVERAGE ARRIVAL DELAY AMONG THE TOP 10 CARRIERS, WHILE SPIRIT AND DELTA HAVE THE LOWEST.



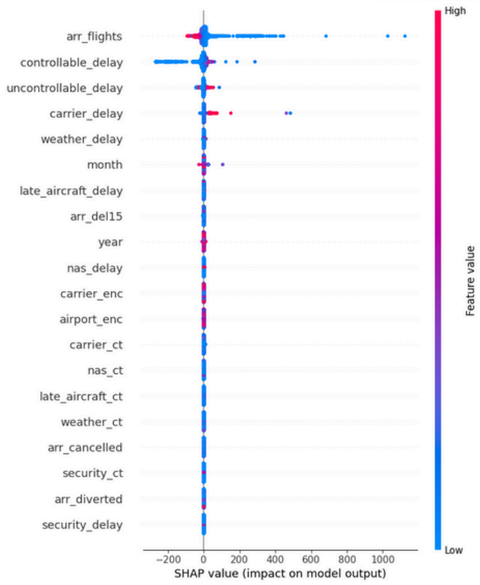
CARRIER AND LATE AIRCRAFT DELAYS CONTRIBUTE THE MOST TO TOTAL DELAYS, WHILE WEATHER, NAS, AND SECURITY DELAYS HAVE A MUCH SMALLER IMPACT.



CORRELATION MATRIX

MODEL DEVELOPMENT & METHODOLOGY

WE USED A TWO-STAGE MODELING APPROACH: (1) CLASSIFICATION TO PREDICT WHETHER A FLIGHT WILL BE DELAYED, AND (2) REGRESSION TO ESTIMATE THE DELAY DURATION. OUR PIPELINE COMBINES DATA PREPROCESSING, FEATURE ENGINEERING, AND MODEL TUNING WITH AN EMPHASIS ON EXPLAINABILITY AND OPERATIONAL RELEVANCE."



REGRESSION TASK – PREDICT DELAY DURATION

- MODEL(S): LINEAR REGRESSION, XGBOOST REGRESSOR, ETC.
- TARGET: DELAY IN MINUTES (FOR DELAYED FLIGHTS ONLY)

SHAP VALUES WERE COMPUTED FOR BOTH CLASSIFICATION AND REGRESSION MODELS TO IDENTIFY THE MOST INFLUENTIAL FEATURES ACROSS ALL PREDICTIONS

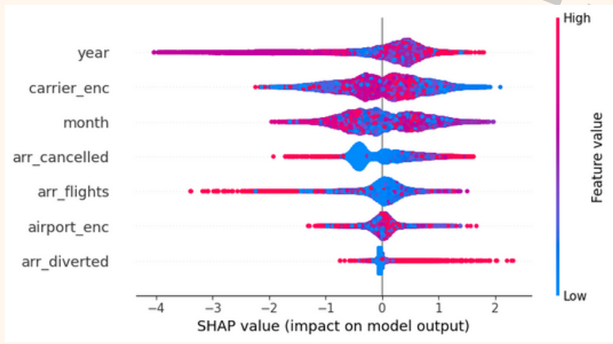
FEATURE ENGINEERING
IMPROVED INPUT QUALITY AND MODEL PERFORMANCE.
AGGREGATED FEATURES (E.G., MEAN DELAY BY ROUTE/AIRPORT/TIME)
CATEGORICAL ENCODING (E.G., ONE-HOT FOR CARRIER, AIRPORT)
DROPPED OR MERGED CORRELATED FEATURES

CLASS IMBALANCE HANDLING (FOR CLASSIFICATION)
DELAYS MAY BE INFREQUENT, LEADING TO A CLASS IMBALANCE.

- TECHNIQUES USED: SMOTE, UNDERSAMPLING, CLASS WEIGHTS
- JUSTIFY METRIC CHOICE (E.G., F2-SCORE TO PENALIZE FALSE NEGATIVES)

REGRESSION MODELS USED
1.LINEAER REGRESSION
2.RANDOM FOREST REGRESSOR
3.XGBOOST REGRESSOR

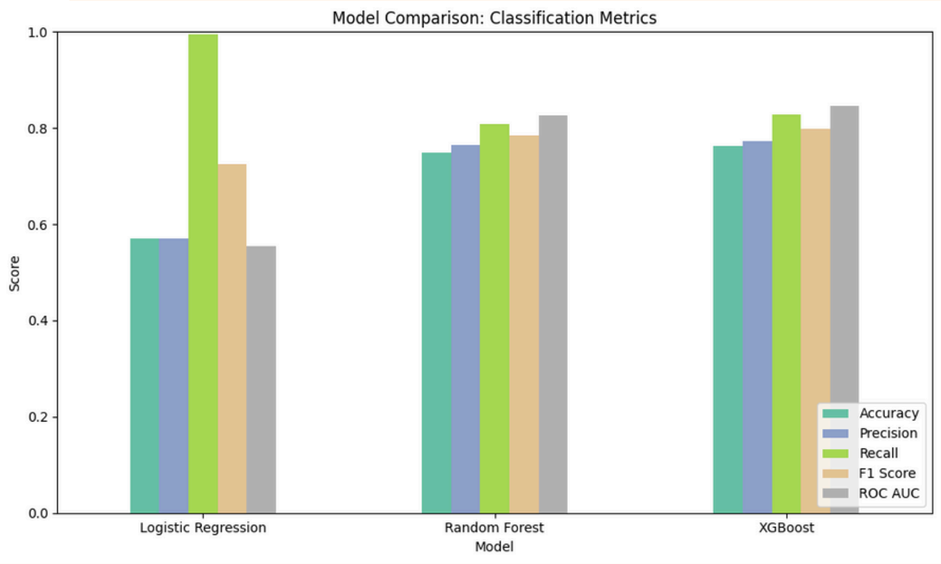
CLASSIFICATION MODELS USED
1.LOGISTIC REGRESSION
2.RANDOM FOREST
3.XGBOOST

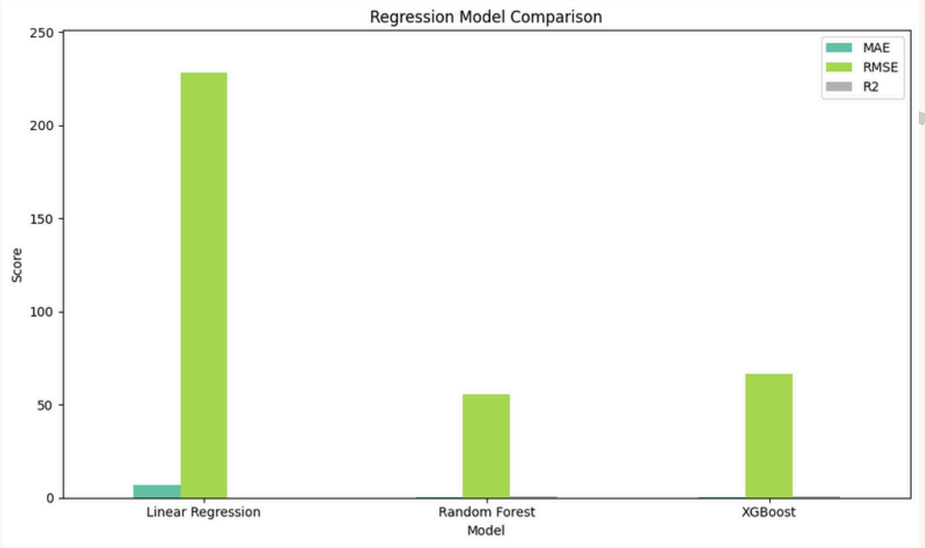


CLASSIFICATION TASK – PREDICT DELAY (YES/NO)

- MODEL(S): LOGISTIC REGRESSION, RANDOM FOREST, XGBOOST, ETC.
- TARGET: BINARY FLAG FOR DELAY (E.G., >15 MINS)

MODEL PERFORMANCE & INTERPRETABILITY





CLASSIFICATION MODEL (PREDICTING DELAY: YES/NO)

KEY METRICS USED:
ACCURACY, ROC AUC
RECALL, F1-SCORE

BEST CLASSIFICATION MODEL

XGBOOST CLASSIFIER
ACCURACY 0.76
PRECISION 0.77
RECALL 0.83
F1 SCORE 0.80
ROC AUC 0.85

REGRESSION MODEL (PREDICTING DELAY DURATION)

KEY METRICS USED:
MAE (MEAN ABSOLUTE ERROR)
RMSE (ROOT MEAN SQUARED ERROR)
R^2 SCORE

BEST REGRESSION MODEL

RANDOM FOREST REGRESSOR
MAE 0.29
RMSE 55.56
R^2 SCORE 0.77

ACTIONABLE RECOMMENDATIONS

BASED ON THE PREDICTIVE INSIGHTS AND SHAP-BASED FEATURE IMPORTANCE, WE IDENTIFIED PRACTICAL STRATEGIES AIRLINES CAN IMPLEMENT TO MITIGATE CONTROLLABLE DELAYS AND ENHANCE OPERATIONAL EFFICIENCY.

A.SCHEDULE OPTIMIZATION

- ADD STRATEGIC BUFFER TIME FOR HIGH-RISK ROUTES OR LATE-DAY FLIGHTS.
- REASSESS TURNAROUND TIMES DURING PEAK CONGESTION PERIODS.

INSIGHT BASIS: DELAYS CLUSTER IN LATE AFTERNOONS AND VARY BY ROUTE.

B. OPERATIONAL EFFICIENCY

- IMPROVE LATE AIRCRAFT MANAGEMENT THROUGH REAL-TIME REALLOCATION.
- IMPLEMENT TIGHTER GATE COORDINATION TO REDUCE GROUND DELAYS.

INSIGHT BASIS: LATE AIRCRAFT AND CARRIER DELAYS ARE THE TOP CONTROLLABLE CAUSES.

C. PROACTIVE COMMUNICATION & CONTINGENCY

- USE MODEL OUTPUTS TO ALERT OPERATIONS TEAMS BEFORE HIGH-RISK FLIGHTS.
- PROVIDE EARLY PASSENGER NOTIFICATIONS TO REDUCE GATE-SIDE DELAYS.

INSIGHT BASIS: PREDICTIVE CLASSIFICATION ENABLES EARLY WARNING SYSTEMS.

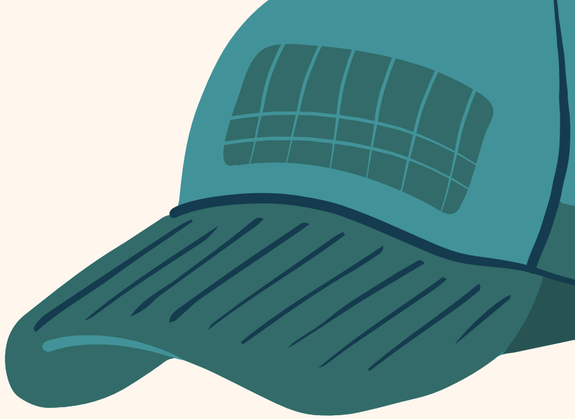
D. DATA-DRIVEN RESOURCE PLANNING

- USE SHAP INSIGHTS TO IDENTIFY ROOT CAUSES AND REDESIGN GROUND WORKFLOWS.
- INVEST IN PREDICTIVE DASHBOARDS TO MONITOR DELAY PROBABILITIES IN REAL TIME.

INSIGHT BASIS: SHAP AND OAI SHOW WHICH VARIABLES CONSISTENTLY IMPACT DELAY LIKELIHOOD.

E. PRIORITIZE CONTROLLABLE DELAYS

BY TARGETING DELAYS ATTRIBUTED TO LATE AIRCRAFT AND CARRIER OPERATIONS — WHICH TOGETHER ACCOUNT FOR OVER XX% OF CONTROLLABLE DISRUPTIONS — AIRLINES CAN MAXIMIZE IMPACT WITH MINIMAL RESOURCE OVERHEAD.



THANK YOU!