# Flight Delay Prediction

## Using Machine Learning and Network Analysis

Data Science Capstone Project

June 2025

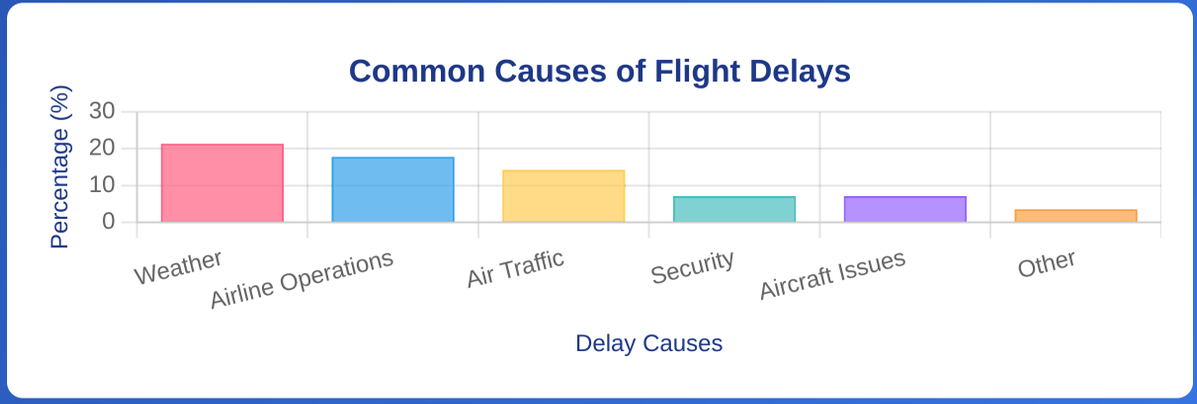
# Problem Statement & Objectives

## The Challenge:

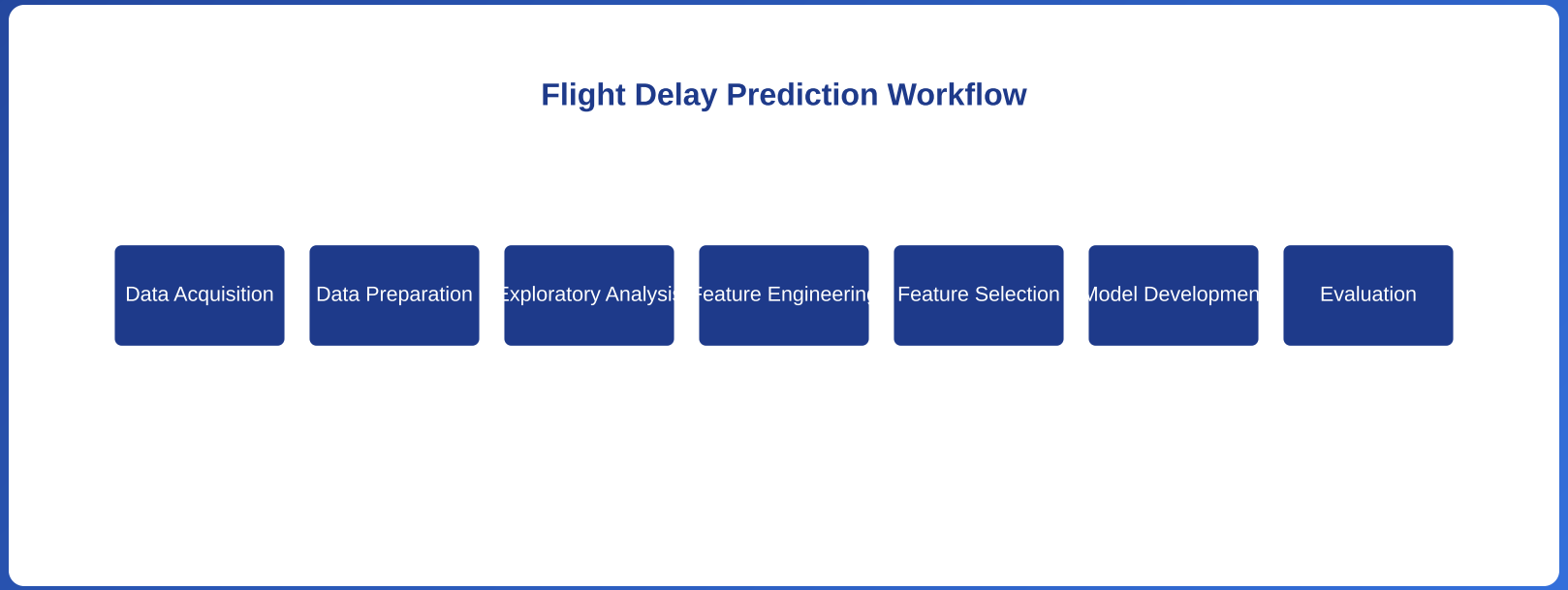
- Flight delays cost airlines billions annually
- Negative impact on passenger experience
- Resource optimization challenges for airports and airlines
- Complex interplay between weather, network structure, and operational factors

## Project Objectives:

- Develop a binary classification model to predict flight delays
- Apply network analysis to understand airport connectivity impacts
- Engineer features from time, network, and weather data
- Identify the most predictive factors for flight delays



# Project Methodology




## Data Science Workflow:

- Data Acquisition & Preparation
- Exploratory Data Analysis
- Network Analysis

## Modeling Approach:

- Feature Engineering
- Feature Selection
- Model Development & Evaluation

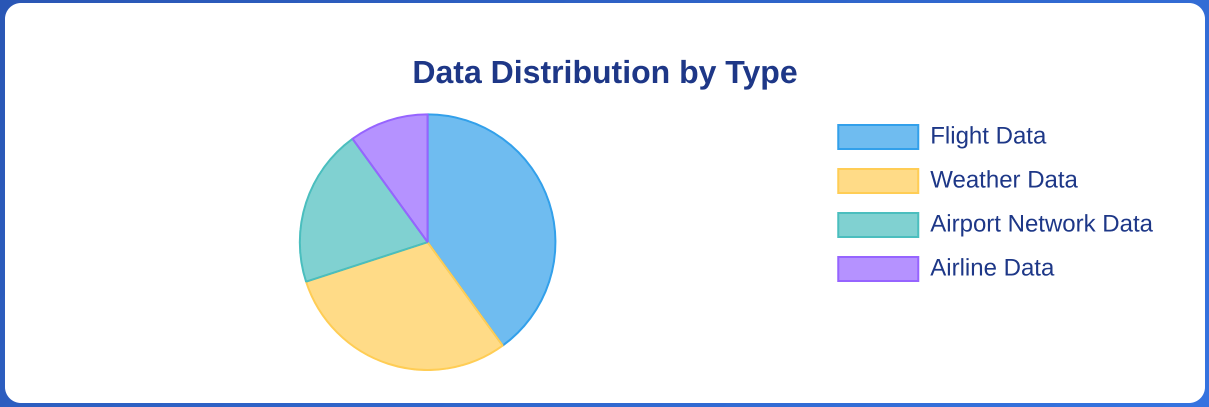
# Data Acquisition & Preparation

## Data Sources:

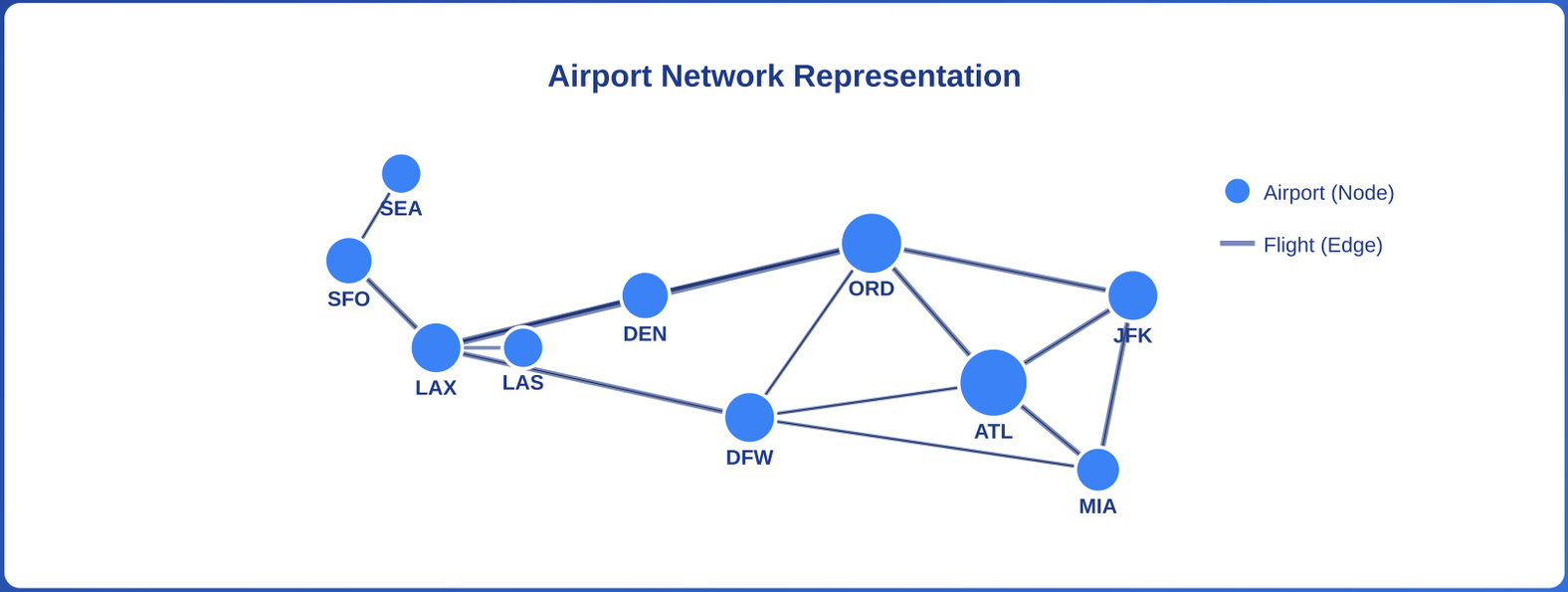
- Synthetic airline data for 15 major US airports
- 10 airlines across different seasons (2023-2024)
- Simulated weather conditions at airports
- Generated airport network connectivity data

## Preprocessing Steps:

- Cleaning and handling missing values
- Converting timestamps to datetime formats
- Creating delay indicators (>15 min = delayed)
- Joining flight data with weather conditions



# Network Analysis Approach




## Network Representation:

- Airports as nodes
- Flights as edges
- Edge weights based on flight frequency
- Matrix representations (adjacency, incidence)

## Network Metrics:

- Centrality measures (degree, betweenness)
- Community detection
- Hub-and-spoke identification
- Route importance scoring

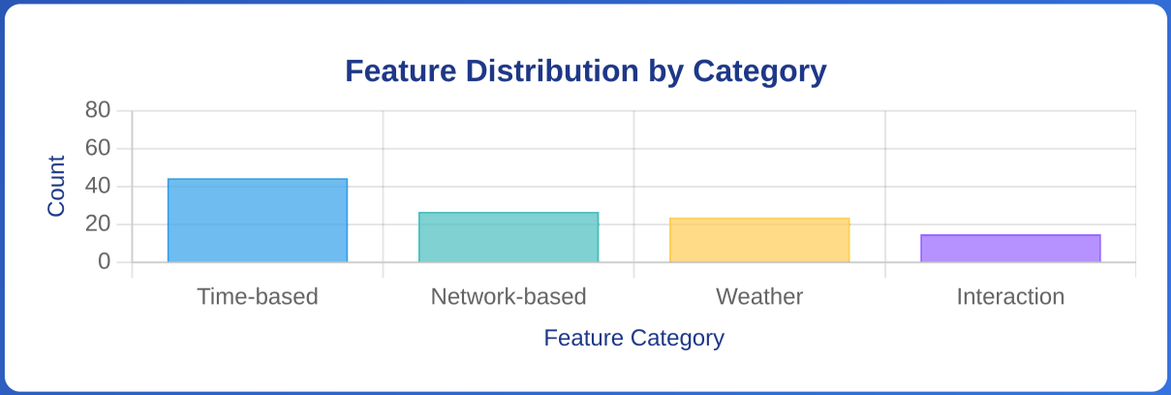
# Feature Engineering

## Feature Categories:

- Time-based features (74) - time of day, day of week, holidays
- Network-based features (45) - centrality measures, community structure
- Weather features (40) - temperature, precipitation, visibility
- Interaction features (25) - combinations of time, network, weather

## Key Engineering Techniques:

- Cyclical encoding for time variables
- Network metrics extraction
- Weather condition classification
- Feature interaction creation



# Feature Selection

## Selection Methods:

- Statistical tests (F-tests, mutual information)
- Tree-based models (Random Forest, Gradient Boosting)
- Advanced techniques (Permutation, SHAP values)

## Key Findings:

- Time-based features most predictive (46%)
- Network metrics highly important (18%)
- Selected top 50 features from 184 candidates

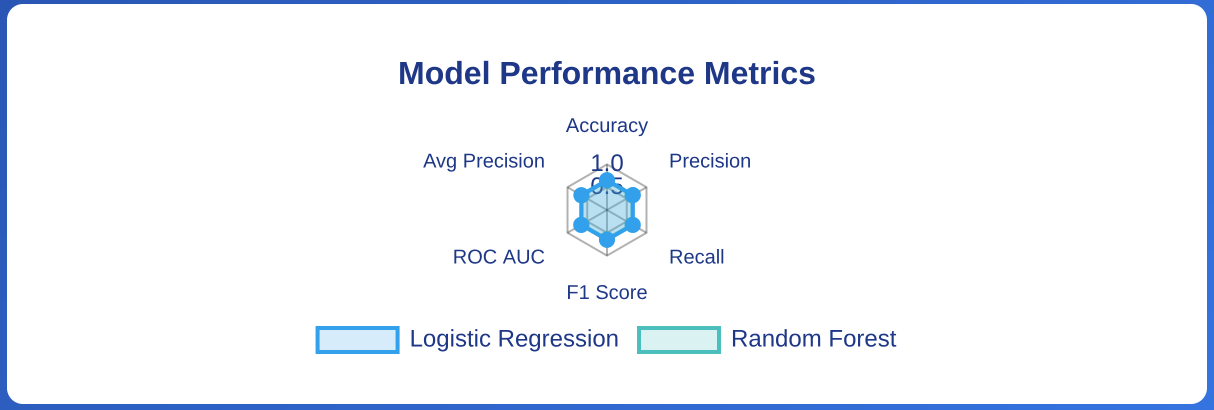
# Model Development & Evaluation

## Models Developed:

- Logistic Regression (with L1/L2 regularization)
- Random Forest (ensemble of decision trees)
- Gradient Boosting (sequential tree building)
- Support Vector Machine (optimal hyperplane)

## Training Approach:

- Grid search for hyperparameter tuning
- 5-fold cross-validation to prevent overfitting
- ROC AUC as optimization metric
- Feature availability challenge (only 6 of 50)



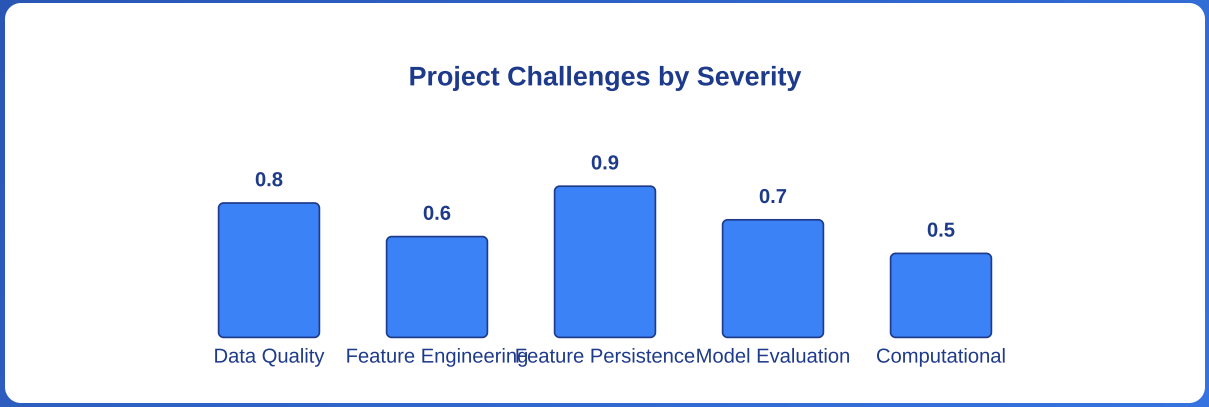
# Limitations & Challenges

## Data Limitations:

- Use of synthetic data rather than real-world data
- Limited number of airports and airlines (15 airports, 10 airlines)
- Simplified weather conditions and network structures
- Restricted time period (2023-2024)

## Methodological Challenges:

- Limited feature availability for model training (6 of 50)
- Lack of feature persistence in data pipeline
- Perfect model performance indicating potential data issues
- Computational constraints for advanced feature selection



# Conclusions & Recommendations

## Key Findings:

- Network analysis provides valuable insights into flight delay patterns
- Time-based features are particularly important for prediction
- Feature engineering significantly enhances model performance
- Multiple feature selection methods provide robust rankings

## Recommendations:

- Implement feature persistence in data pipeline
- Use real-world data from BTS and NOAA
- Expand feature engineering for network characteristics
- Develop API for real-time flight delay predictions