# Predicting Flight Delays Using Historical Airline and Weather Data

## Capstone Project Proposal Summary

### Problem Statement & Context

The airline industry faces significant challenges related to flight delays, resulting in billions of dollars in annual economic losses, customer dissatisfaction, and operational inefficiencies. Despite technological advancements, predicting delays remains complex due to numerous contributing factors. This project aims to develop a robust predictive model that forecasts flight delays by leveraging historical airline performance data and weather information, focusing on advanced time series analysis techniques.

Flight delays affect approximately 20% of all U.S. flights and cost airlines an estimated $8.3 billion annually. Weather conditions account for approximately 70% of all delays, but their impact varies across airports and aircraft types. Other factors include air traffic congestion, mechanical issues, airline scheduling practices, and seasonal patterns. Recent advancements in data science, particularly in time series analysis, combined with comprehensive historical data from the Bureau of Transportation Statistics and NOAA, provide an unprecedented opportunity to develop sophisticated predictive models.

### Criteria for Success

Success will be measured through both technical performance and practical utility metrics:

* Predictive accuracy: MAE of less than 15 minutes for delay predictions and at least 80% accuracy in classifying delayed versus on-time flights
* Actionable insights with 6-12 hour lead time for implementing mitigation strategies
* Model robustness across different airports, airlines, seasons, and weather conditions
* Interpretability of factors driving flight delays
* Potential for integration into existing airline and airport operations.

### Scope of Solution & Methodology

The solution will employ time series analysis to capture temporal patterns in flight operations, enhanced with weather data integration. The approach will explore various methodologies including ARIMA, SARIMA, and LSTM networks to capture complex temporal dependencies. Feature engineering will transform raw data into meaningful predictors, including lag features and derived variables capturing weather-operational interactions. The solution will include a comprehensive data pipeline for acquiring, cleaning, and integrating data from multiple sources, while prioritizing both prediction accuracy and interpretability.

### Constraints

Key constraints include:

* Data quality and availability challenges, including inconsistencies and missing values
* Computational resource limitations when working with large-scale historical datasets
* The inherent complexity and stochasticity of the aviation system
* Project timeframe constraints limiting exploration depth
* Implementation constraints related to industry regulations and stakeholder technical capabilities.

### Stakeholders

Primary stakeholders include:

* Airlines (American, Delta, United) seeking to optimize scheduling and resource allocation
* Airport authorities managing infrastructure and ground operations
* Air traffic control organizations (FAA) responsible for airspace utilization
* Passengers requiring reliable flight information
* Travel industry partners (hotels, car rentals) affected by flight delays
* Academic and research institutions interested in transportation analytics

### Data Sources

The project will utilize two primary data sources:

1. BTS Airline On-Time Performance Database: Contains over 200 million flight records since 1987, including scheduled/actual times, airports, carriers, and delay categorizations. Accessible via <https://www.transtats.bts.gov/Tables.asp?DB_ID=120>.
2. NOAA Climate Data: Provides meteorological information including temperature, precipitation, wind, visibility, and barometric pressure from thousands of stations. Accessible via [https://www.ncdc.noaa.gov/cdo-web/.](https://www.ncdc.noaa.gov/cdo-web/)

Integration challenges include temporal and spatial alignment of weather observations with flight operations, data cleaning, and handling missing values. Supplementary sources may include FAA's Aviation System Performance Metrics database and holiday/event calendars.

### Deliverables

The project will produce:

* A comprehensive Python codebase implementing the entire data pipeline and modeling workflow
* A technical report detailing methodological approach, findings, and model performance
* An executive summary with key insights and recommendations
* A presentation slide deck for both technical and non-technical audiences

These deliverables will demonstrate the application of advanced time series analysis to a complex real-world problem, showcasing both technical sophistication and practical utility in addressing a significant challenge in the aviation industry.