Improving Customer Experience

via Flight Delay Predictions

Team 1-1

Yoni Nackash, Nicholas Lin, Asma Farooq,
Nicholas Bermingham, Mian Haseeb

Agenda

- Team Details & Introduction
- EDA
- Feature Engineering and Top Features
- Model Pipeline
- Model Baseline
- Model Expansion
- Conclusion & Next Steps

Team

Yoni Nackash
yoninackash@berkeley.edu

Nicholas Lin
nicholaslin@berkeley.edu

Asma Farooq
asmafarooq@berkeley.edu

Nicholas Bermingham
nbermingham@berkeley.edu

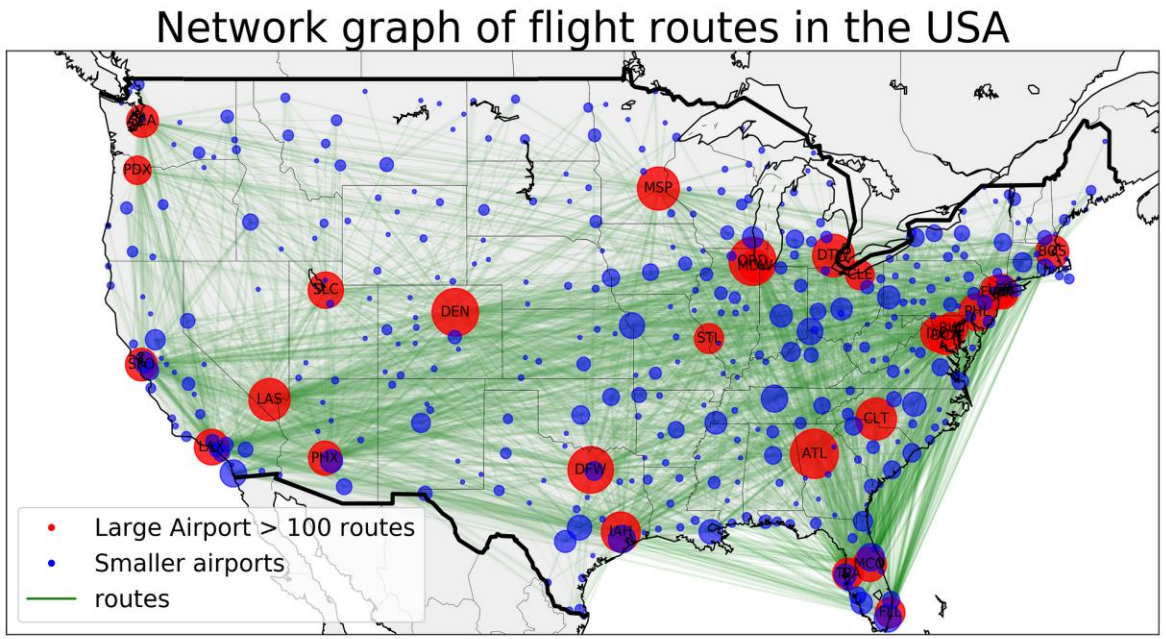
Mian Haseeb
mianh1@berkeley.edu

Abstract

- According to the Transporation Security Agency (TSA), air travel demand and flight delays will continue to increase in parallel
- Leveraged machine learning models (Random Forest, XGBoost, Neural Networks) to predict flight delays using historical data from 2015 to 2019, including weather conditions.
- Random Forest model improved RMSE by 18%, but hyperparameter tuning with Optuna underperformed due to insufficient trials.
- Accurate prediction of longer delays remains difficult; models performed better for short delays.
- In future we recommend refining models, adding new features, and utilizing more computational resources for enhanced predictive accuracy.

Project Description & Motivation

- BAC (Berkeley Airline Company) hired our Data Science Team to provide flight delay prediction modelling
- Analyzing flight data between 2015-2019, on average 15% of flights were delayed
- Our objective is to predict flight delays two hours prior to departure, in order to improve the customer experience and mitigate extraneous airline costs



Data Description

Flight Data (US Department of Transportation)

- Contains flight data departing from all major US airports. We will focus on flights between the years of 2015 - 2019
- This dataset contains important flight information such as date, airline, origin, destination, departure time, arrival time, and various fields describing delays

Weather Data (National Oceanic and Atmospheric Administration Repository)

- Contains weather readings from stations throughout the US. We will focus on readings between the years 2015 and 2019, as well as readings in relation to the airports in our flight data
- Contains pertinent weather data such as temperature, visibility, sky coverage, and wind

Station Data

- Contains metadata about each major airport that we ewill use to join the flight and weather data. Includes 2,261 distinct weather stations

Airport Data

- Contains 54,380 unique airports that fall into these categories: large_airport, medium_airport, small_airport, balloonport, seaplane_base, and heliport,

EDA – Flight Data

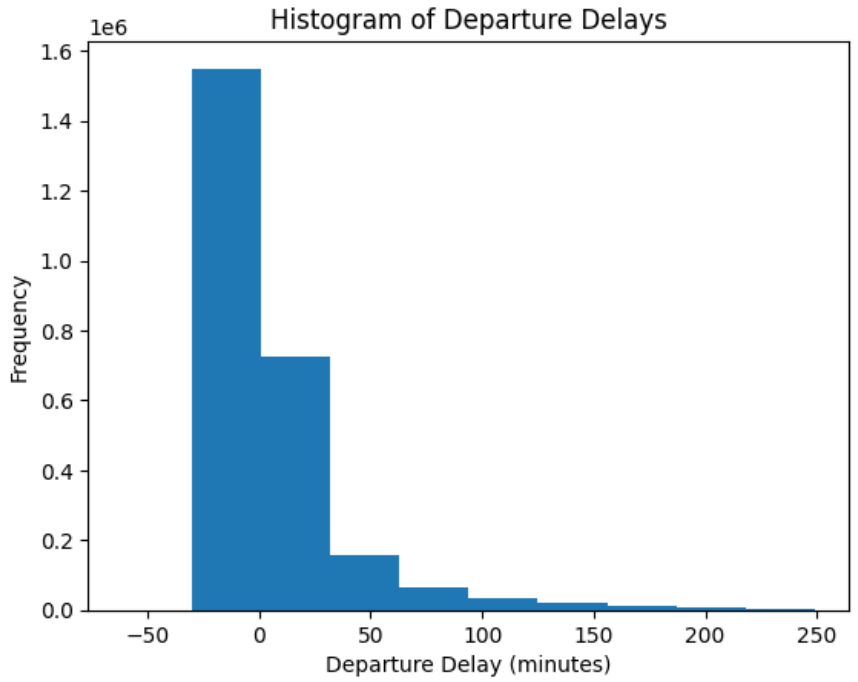


Figure 1:
Outliers of delays beyond 250 minutes trimmed

Figure 2:
Flight Delay % On Non- Holidays vs Holidays

EDA – Flight Data

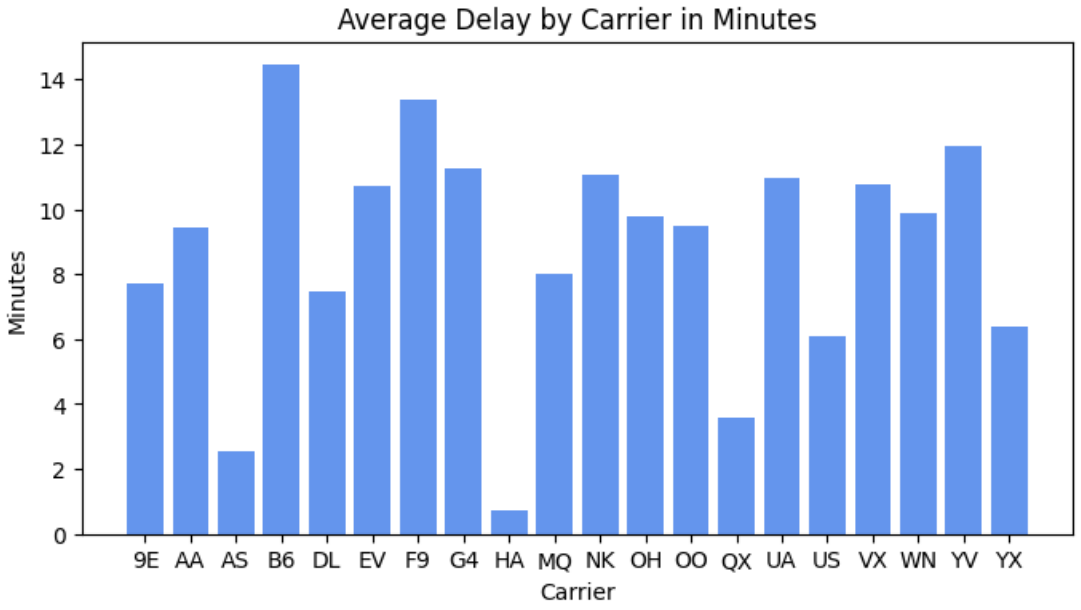


Figure 3:
Average Delay by Carrier In Minutes

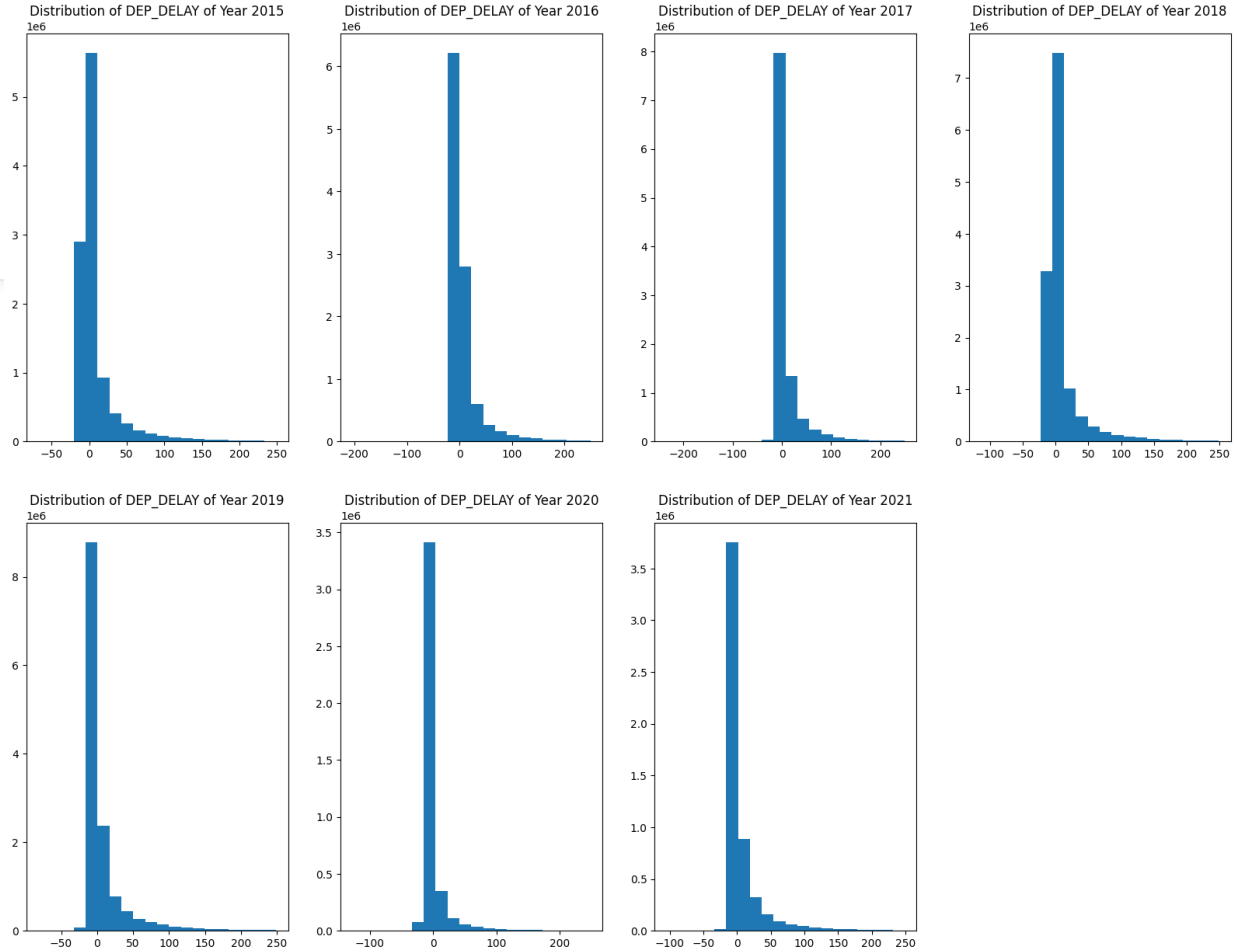
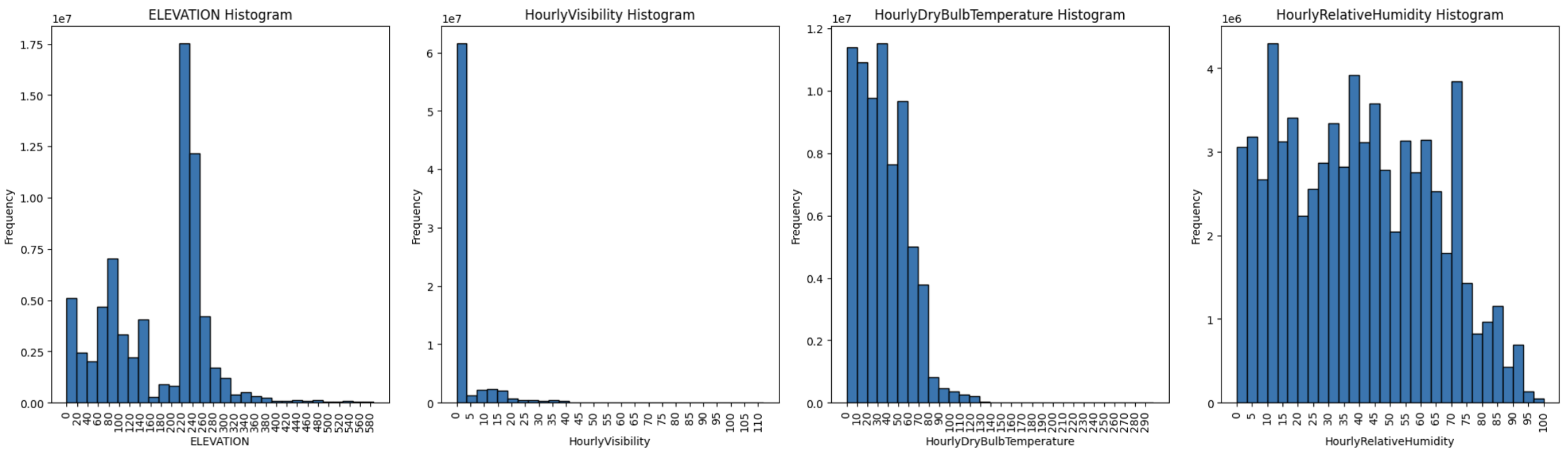
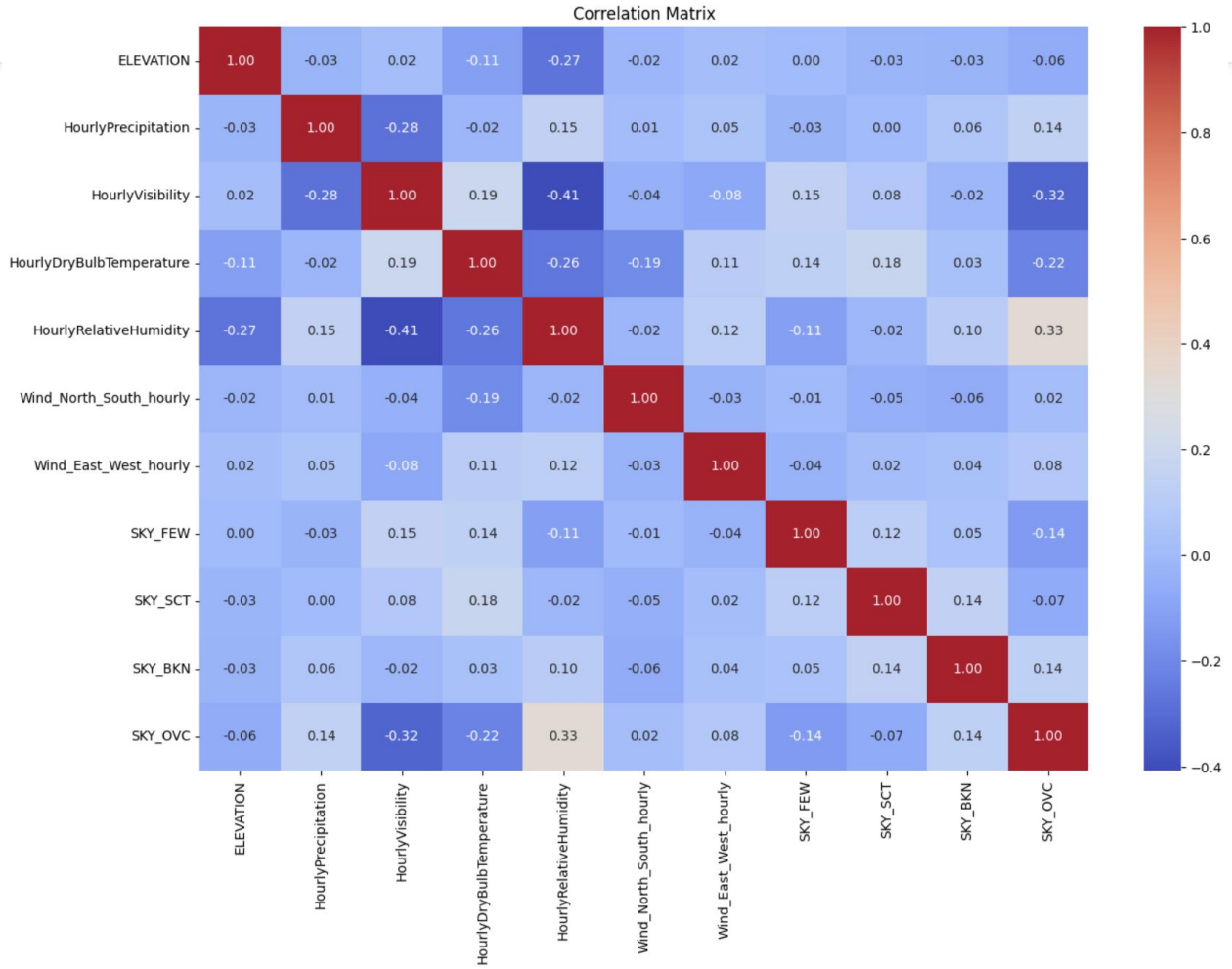


Figure 4:
Delay Distribution by Years

EDA - Weather Data

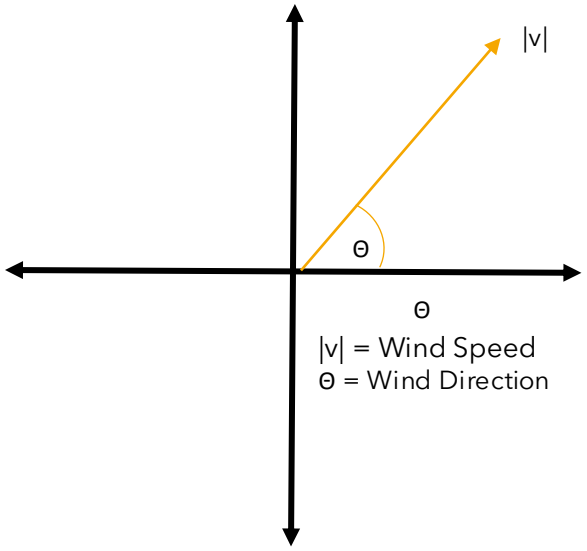


EDA – Correlation Matrix

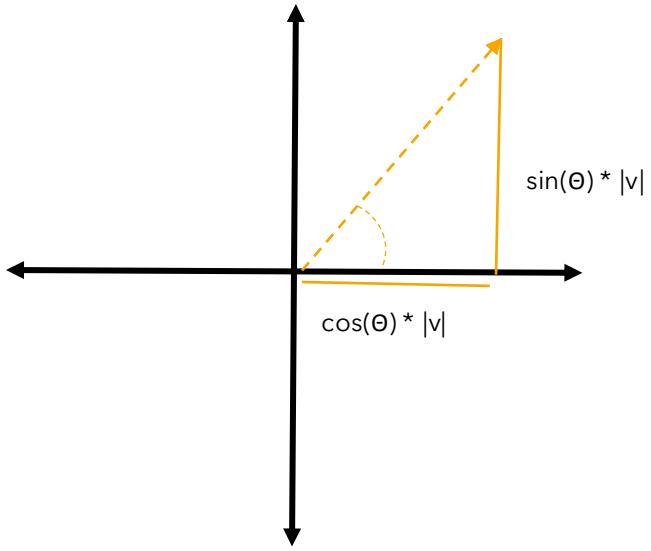


Feature Engineering Highlights: Weather

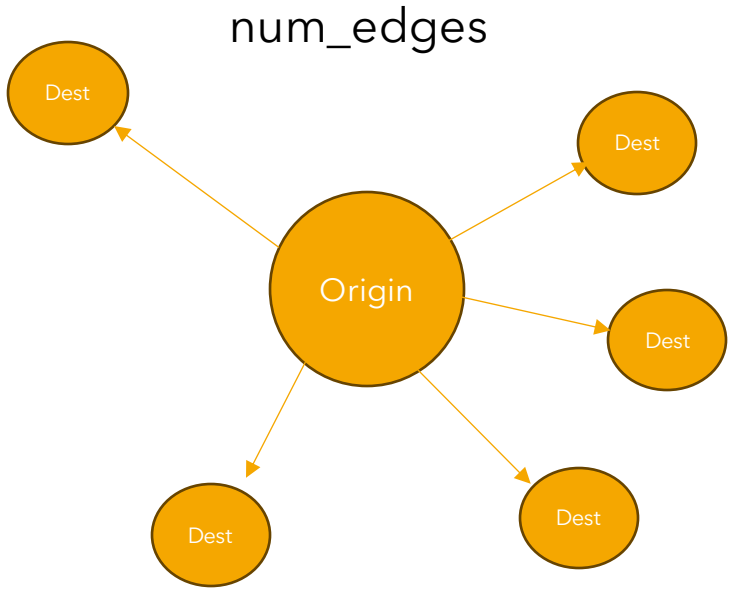
Original data: Wind Speed and Direction



Engineered Features: Wind Components (North-South & East-West)



Feature Engineering Highlights: Airports



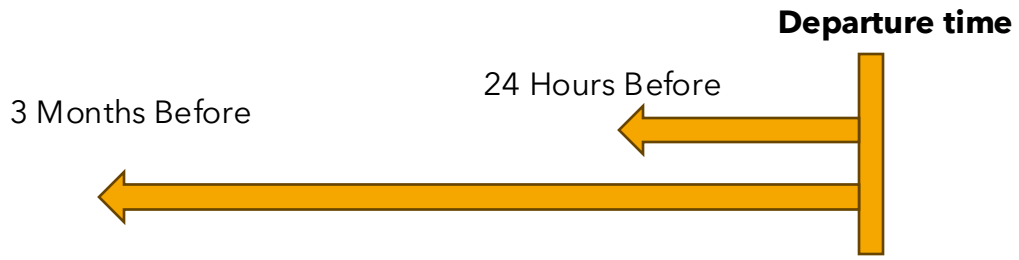
Aggregations

- Total flights 2-8 hours before
- Number of delays 2-8 hours before
- Percent Delays 2-8 hours before

Feature Engineering Highlights: Tail Num

Delay Metrics:

- Average delay within 24 hours and 3 months.
- Percentage of delays above 15 minutes.



Features Selected

Weather Features:

- HourlyPrecipitation, HourlyRelativeHumidity, HourlyVisibility, HourlyDryBulbTemperature, Wind_North_South_hourly, Wind_East_West_hourly, Season, SKY_FEW, SKY_SCT, SKY_BKN, SKY_OVC

Date Features:

- DAY_OF_WEEK, weekend, is_holiday, within_3_days_of_holiday, time_of_day

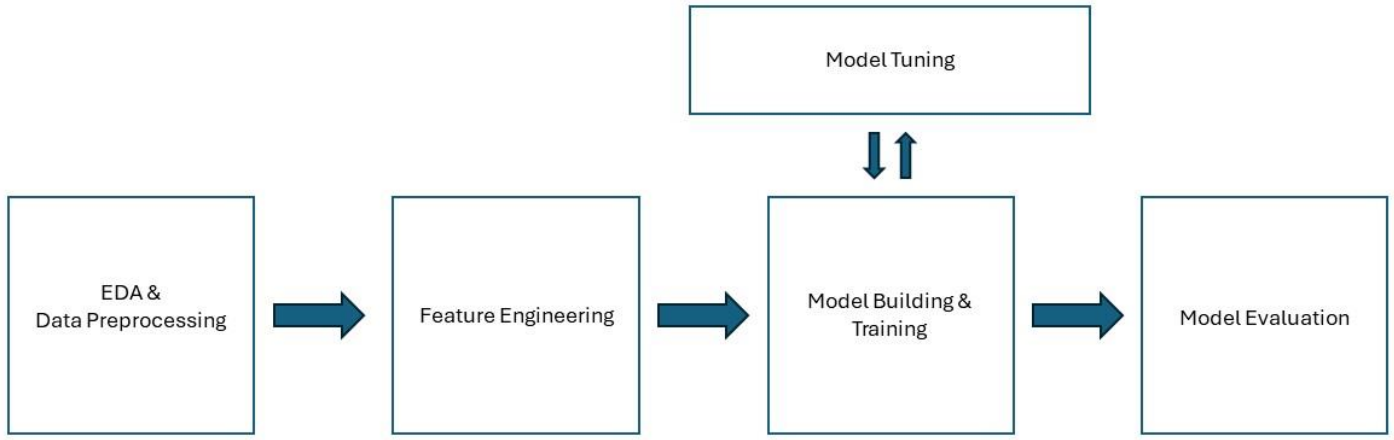
Flight + Station Features:

- OP_UNIQUE_CARRIER, ORIGIN, DEST, origin_type, avg_delay_24hrs, avg_delay_3months, delay_24hrs_pct, delay_3months_pct, num_edges, pct_flights_2_8hrs_delayed, count_flights_2_8hrs

Machine Learning Pipeline

Phase 3 Models Explored: XGBoost Regressor, Random Forest, Neural Network Regressors

Data Split: Training & Cross Validation Set (2015-2018), Testing Set (2019)



Model Baseline

(Used the mean of "DEP_DELAY" to evaluate against "DEP_DELAY")

Training RMSE: 29.69

Training MAE: 17.19

Training R2: 0.00

Test RMSE for delays > 60: 114.31

Test R^2 for delays > 60: -5.28718

Test RMSE for delays <= 60: 14.70

Test R^2 for delays <= 60: -0.212175

Results

Test RMSE: 31.52

Test MAE: 18.14

Test R^2 : 0.00

Ensemble Models – *XG Boost*

Training Time: 57 mins

Training RMSE: 28.092

Training MAE: 15.328

Training R2: .1305

Test RMSE for delays > 60: 102.613

Test MAE for delays > 60: 91.367

Test R^2 for delays > 60: 4.282

Test RMSE for delays <= 60: 14.153

Test MAE for delays < 60: 10.713

Test R^2 for delays <= 60: -0.187

Results

Test RMSE: 27.647

Test MAE: 15.133

Test R^2 : .1331

Ensemble Models - *Random Forest*

Training RMSE: 28.4993

Training MAE: 15.9285

Training R2: 0.11907

Test RMSE for delays > 60: 104.7744

Test R^2 for delays > 60: -4.5471

Test RMSE for delays <= 60: 13.2916

Test R^2 for delays <= 60: -0.1072

Results

Test RMSE: 25.7832

Test MAE: 14.3692

Test R^2 : 0.1156

Ensemble Models - *Random Forest Hyperparameter Tuning*

Training RMSE: 28.5819

Training MAE: 15.9727

Training R2: 0.1139

Results

Test RMSE: 25.8873

Test MAE: 14.4489

Test R^2 : 0.1084

Proof of Concept: PyTorch Neural Network

- Due to **memory constraints** of PyTorch, the data used is only from January, 2015.
    - **Training**: 1/1/2015 - 1/20/2015
    - **Test**: 1/21/2015 - 1/31/2015

Training RMSE: 27.4

Layers

- Linear (819, 128)
- Dropout
- Linear (128, 64)
- Dropout
- Linear (64, 32)
- Linear (32,1)

Results

Test RMSE: 24.0

Conclusions and Future Improvements

- **Best Performing Model (Full Dataset)**: Random Forest (18% improvement from baseline)
- **Number of Features**: 28
- **Top 10 Features**: OP Unique Carrier, Origin, Destination, Origin Type, Day Of Week, Season, Time Of Day, Elevation, Hourly Precipitation, Hourly Visibility
- **Hyper Parameters**: WIP

References

1. https://www.cnbc.com/2024/06/28/record-summer-air-travel-is-starting-and-flight-hassles-and-delays.html