Predicting Flight Delay

Team 1-2: Alec, Jian, Patrick, Trisha

About the Team

Alec Naidoo

San Francisco, CA

Jian Wang

Vancouver, BC

Patrick Yim

San Francisco, CA

Trisha Sanghal

Austin, TX

Presentation Outline

Abstract

Accurate flight delay predictions can be vital to the success of airlines for multiple reasons: (1) airlines can optimize their resource and personnel allocation to minimize disruption to passengers, (2) airlines can make more informed long-term decisions about infrastructure improvements, and (3) findings from previous flight delay responses can help airlines advance research and development across the broader aviation industry. To help contribute to the success of airlines, our project's aim is to produce a binary classification model that predicts whether or not a flight will experience arrival delay, defined as a 15-minute or greater difference between the planned and actual arrival times. We believe that the binary classification model will produce predictions that are easy to interpret and action on (airlines only need to know whether the delay exceeds the 15-minute threshold to decide whether to implement targeted responses), and will be better suited than a regression model to handle data imbalances in arrival delay. To accomplish this goal, we will leverage the following three datasets: flight on-time performance data from the U.S. Department of Transportation, weather data corresponding to origin and destination airports from the National Oceanic and Atmospheric Administration Repository, and metadata about airports in the U.S. from the U.S. Department of Transportation. These datasets contain data from 2015 to 2019 for flights within the United States.

Our first baseline model always predicts the majority class of no delay, and has an F1 score of 71.96% on the test set (last quarter of the 1-year 2015 OTPW data). Our second baseline model predicts delay or no delay at random, and has an F1 score of 55.14% on the test set. We selected these models as our baselines because they are simple to implement and understand, and they provide us with a reasonable minimum performance level (if our model doesn't outperform the baselines, it suggests that our model isn't capturing useful patterns from the data). We also implemented logistic classification and decision tree classification models. A few of the most important features we used as inputs to these models are 'airline_performance', 'DISTANCE', 'CRS_ELAPSED_TIME', 'CRS_DEP_TIME', 'HourlyWetBulbTemperature', and 'HourlyWindSpeed'.

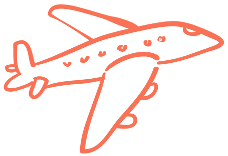
Some of our next steps involve implementing dimensionality reduction (through PCA and regularization), and building and testing a neural network classification model.

Project Description: Data Description

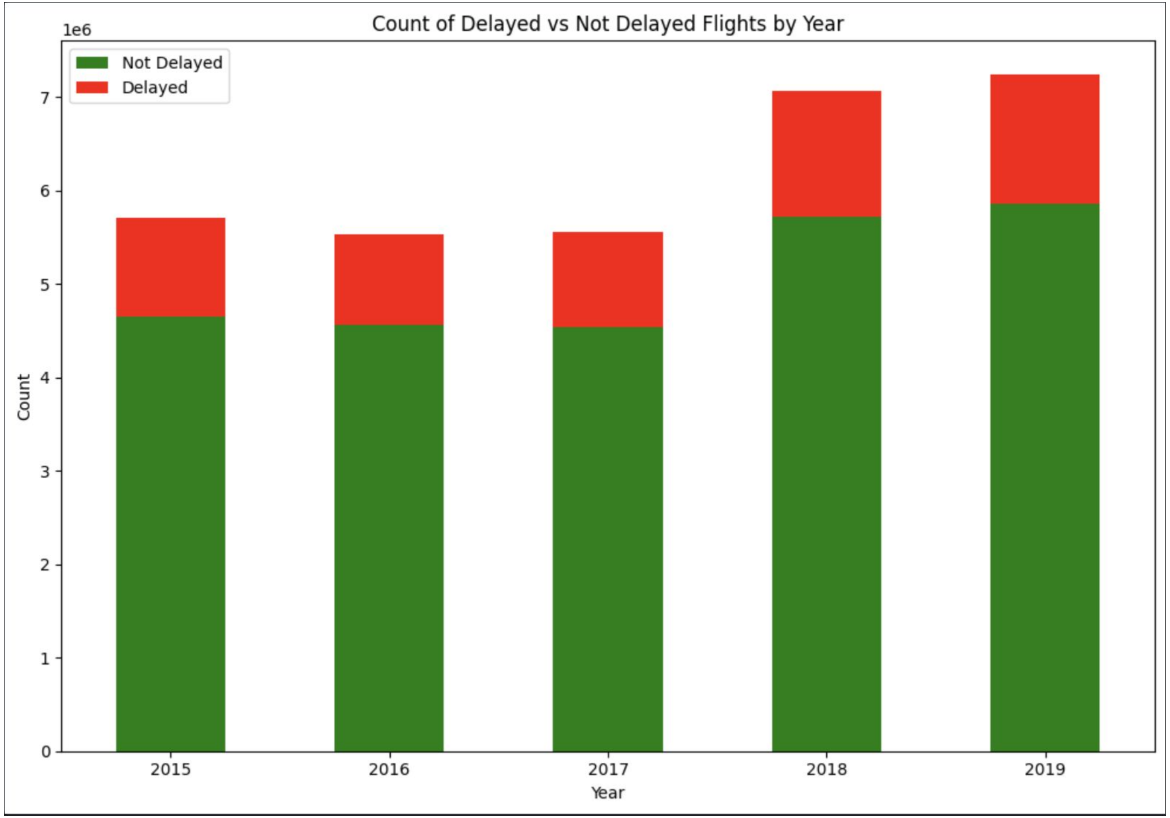
Dataset	Source	Timeframe Available	Description	Dimensions
Flights	TranStats Data Collection from the U.S. Department of Transportation	2015 to 2019	Contains flight on-time performance data within the U.S.	31,746,841 x 109
Weather	National Oceanic and Atmospheric Administration Repository	2015 to 2019	Contains weather data corresponding to origin and destination airports at departure and arrival times.	630,904,436 x 124
Stations	U.S. Department of Transportation	Last Updated 2024	Contains metadata about airports in the U.S.	18,097 x 12
OTPW (Ontime Performance of Flights and Weather)	261 Instructors	2015 to 2019	Contains joined data from the flights, weather, and stations tables.	31673119 x 214

Project Description: Our Task

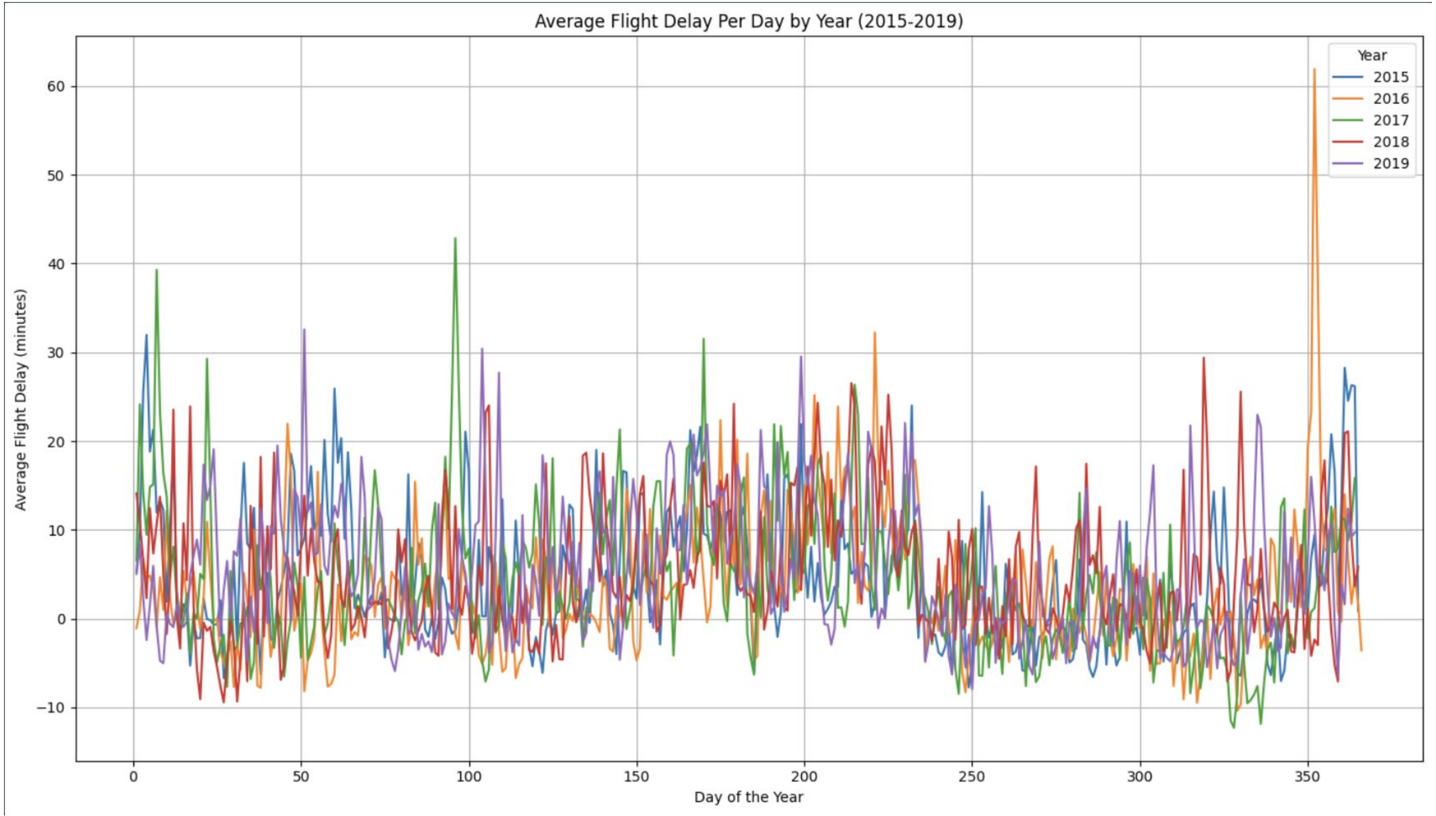
In this project, we aim to develop a **binary classification model** that can predict whether or not a flight within the United States will experience **arrival delay** given information about the flight, airport, and weather conditions.



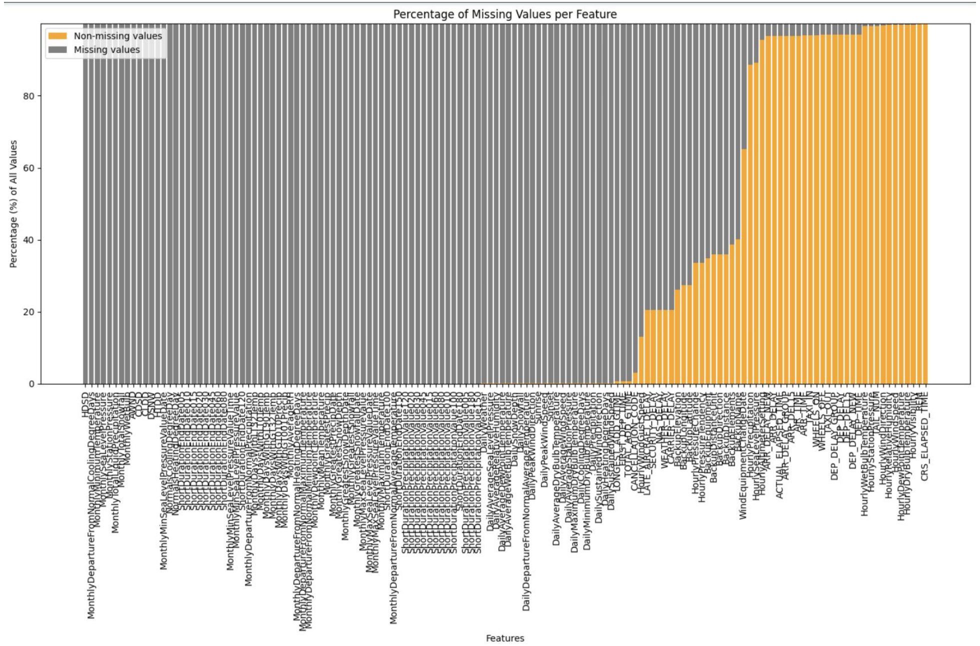
Target Variable: ARR_DEL15



EDA: Average Arrival Delay Throughout the Year

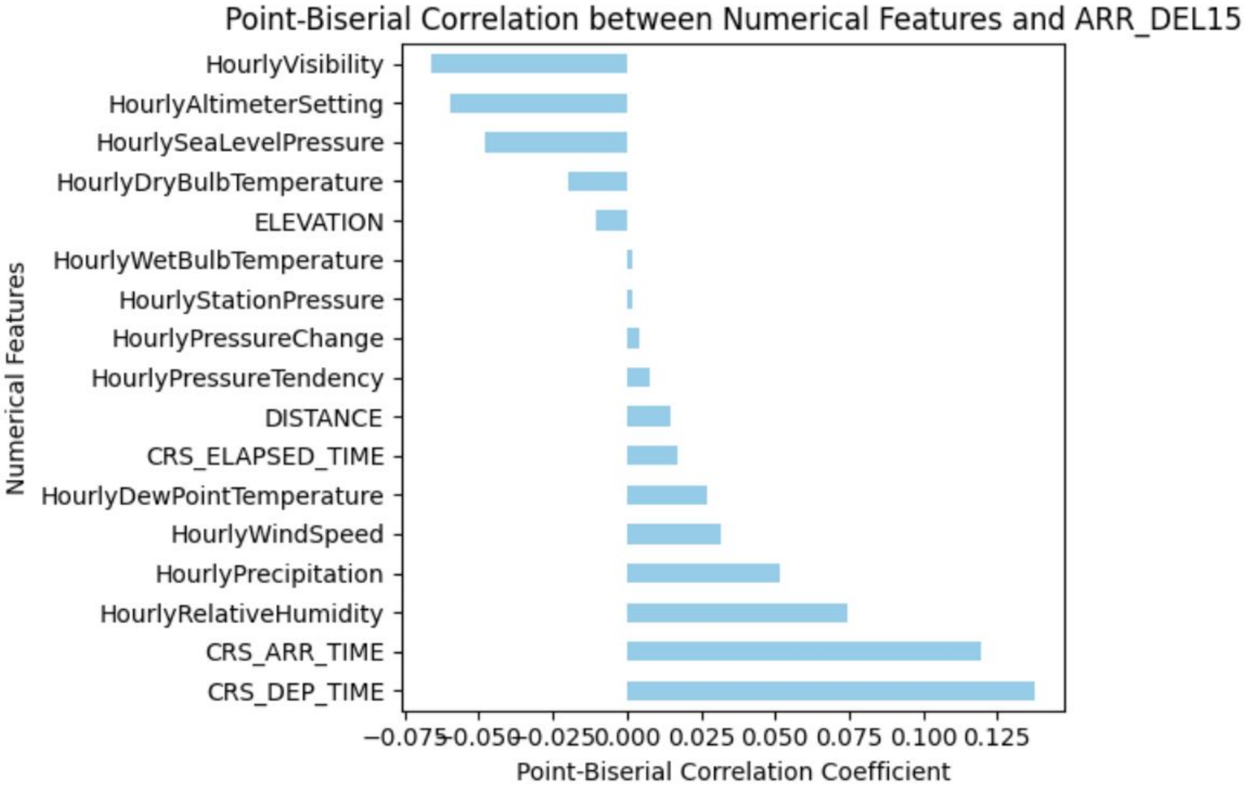


EDA: Missing Value Analysis



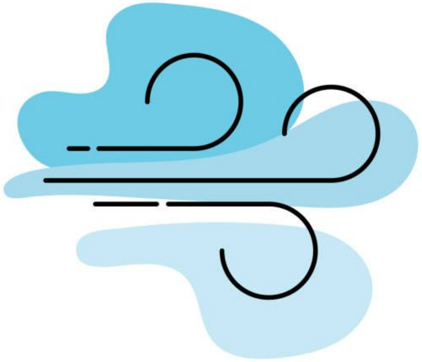
- **Numeric Columns:** Cast to double and impute missing values with the mean of each column.
- **Non-Numeric Columns:** Impute with placeholders or the mode (e.g., WindEquipmentChangeDate filled with 1900-01-01).
- **Specific Columns:** Use targeted strategies like mode for HourlyPressureTendency and placeholders for HourlyWindDirection.
- **Drop Columns with > 80% Nulls**

Point Biserial Correlation Coefficients

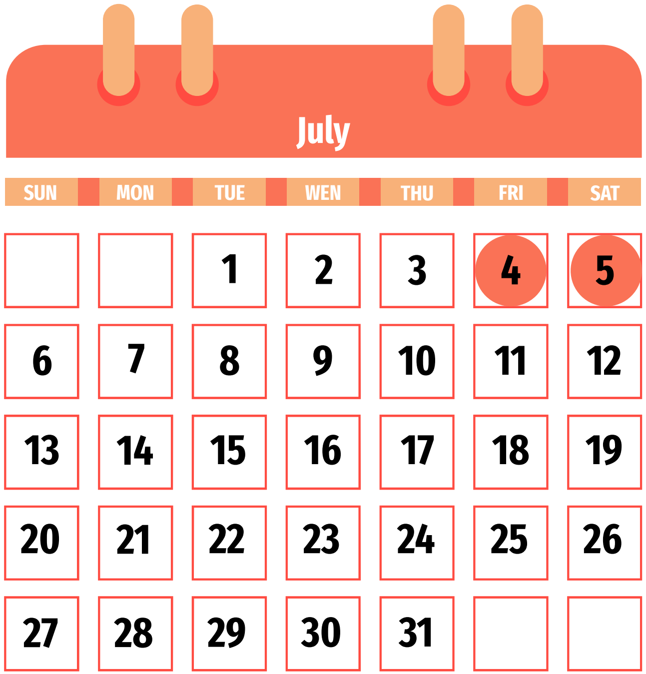


Feature Engineering: Wind Variable/Direction

Variable Wind	Count
1	250051
0	5456837

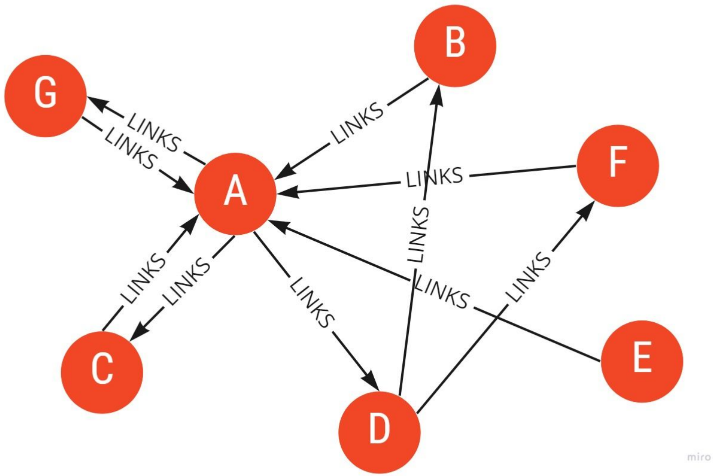


Feature Engineering: Holiday/Weekend



Feature Engineering: Pagerank

Origin Pagerank	Destination Pagerank
4.136412399	0.480389757
4.136412399	0.480389757
4.136412399	0.480389757
4.136412399	0.480389757
7.09744475	0.480389757
7.09744475	0.480389757
4.695157215	0.480389757
4.695157215	0.480389757
4.695157215	0.480389757



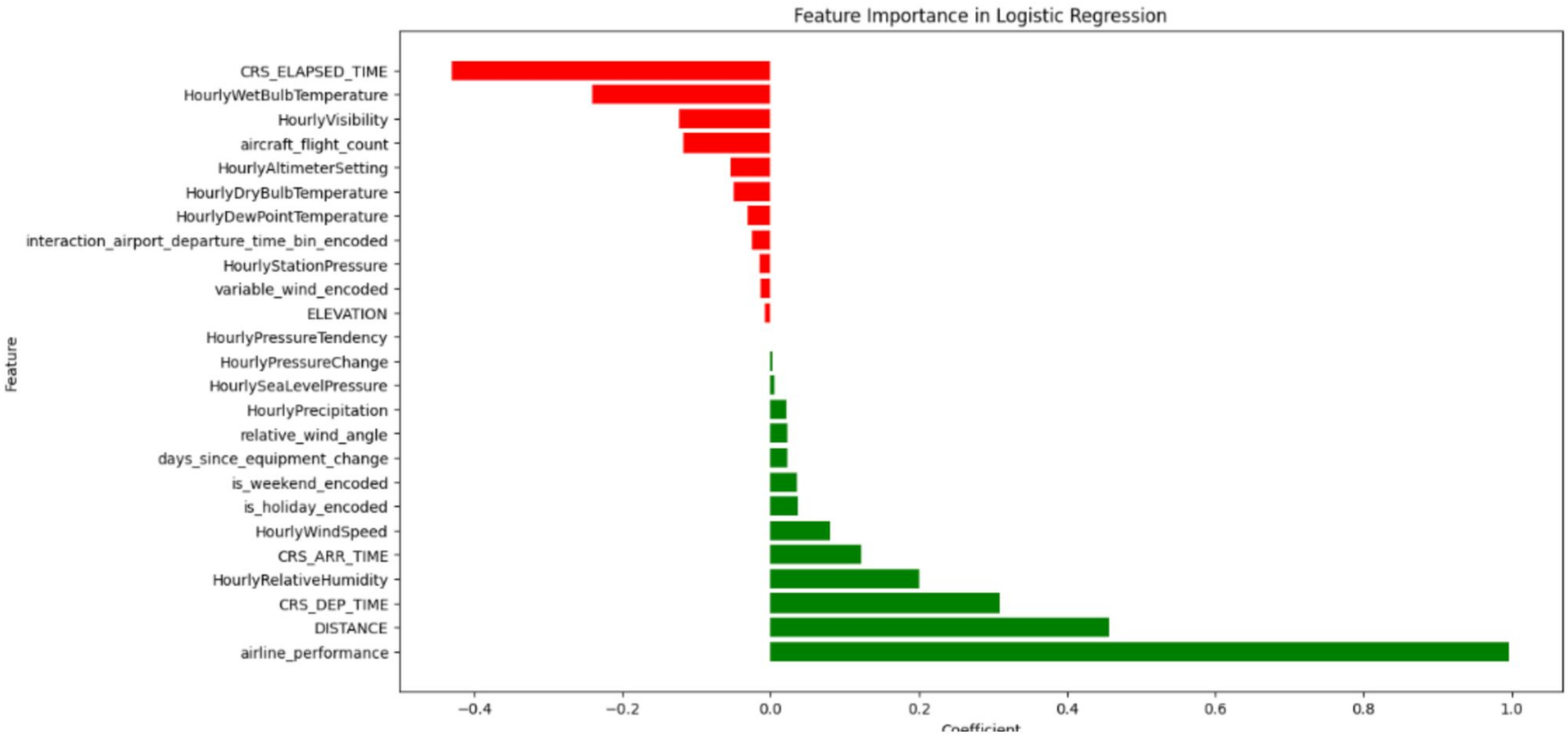
Feature Engineering: Derived Features

Feature Name	Description	Implementation	Rationale
Interactive Term (Airport x Departure Time)	Create an interaction term between the airport and the scheduled departure time.	`interaction_airport_departure_time = airport * scheduled_departure_time`	Different airports might have varying levels of congestion at different times of day.
Wind Resistance	Create Wind Direction Feature to see if there was variable wind.	`Variable_wind = 1 if wind direction is marked variable else 0`	Flights can be impacted by wind causing delays.
Holiday Indicator	Determine if the flight date falls on or near a holiday.	`is_holiday = 1 if flight_date in holiday_dates else 0`	Flights around holidays may have different delay patterns due to higher travel volumes.
Weekend Indicator	Determine if the flight date falls on a weekend.	`is_holiday = 1 if flight_date in weekend_dates else 0`	Flights around weekends may have different delay patterns due to higher travel volumes.
Days Since Equipment Change	Calculate the number of days since the last equipment change.	`Days_since_equipment_change = flight_date - wind_equipment_change_date`	Older equipment might be more prone to causing delays.
Aircraft Flight Count	Calculate the flight count of the aircraft.	Count the number of flights for each aircraft and update the count for each row.	Older aircrafts might be more prone to mechanical issues causing delays.
Airline Performance	Capture the historical performance of airlines (7 days).	Average delay time window, on-time performance metrics.	Airlines with better performance records are likely to have less delays.

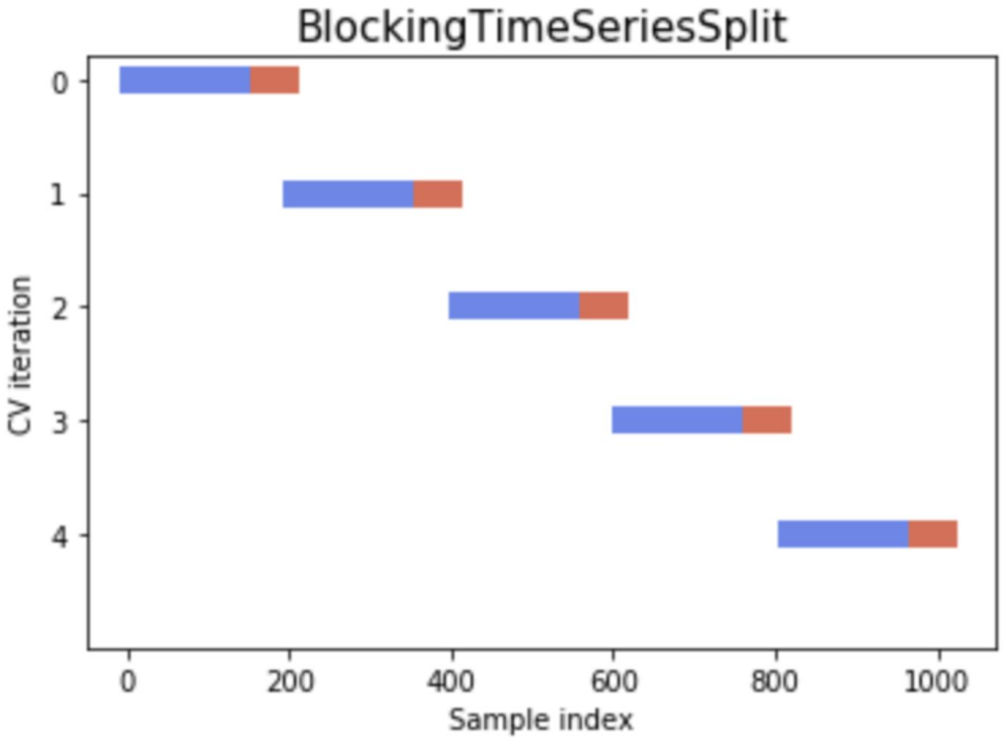
Feature Engineering: Final Features

	Column Name	Missing Values Count
0	CRS_DEP_TIME	0
14	HourlyVisibility	0
25	is_weekend	0
24	is_holiday	0
23	variable_wind	0
22	dest_airport_pagerank	0
21	origin_airport_pagerank	0
20	airline_performance	0
19	aircraft_flight_count	0
18	days_since_equipment_change	0
17	relative_wind_angle	0
16	HourlyWindSpeed	0
15	HourlyWetBulbTemperature	0
13	HourlyStationPressure	0
1	CRS_ARR_TIME	0
12	HourlySeaLevelPressure	0
11	HourlyRelativeHumidity	0
10	HourlyPressureTendency	0
9	HourlyPressureChange	0
8	HourlyPrecipitation	0
7	HourlyDryBulbTemperature	0
6	HourlyDewPointTemperature	0
5	HourlyAltimeterSetting	0
4	ELEVATION	0
3	DISTANCE	0
2	CRS_ELAPSED_TIME	0
26	ARR_DEL15	0

Feature Engineering: Top Features



Modeling Pipelines: Cross Validation



Modeling Pipelines: Models Explored

Algorithm	Input Features	Loss Function	Evaluation Metrics
Majority Class Baseline	None	None	- Precision - Recall - Accuracy - **F1 Score**
Random Class Baseline	None	None	- Precision - Recall - Accuracy - **F1 Score**
Logistic Classification	- Numerical: 21 - Categorical: 4 - Derived: 4	Log Loss	- Precision - Recall - Accuracy - **F1 Score**
Random Forest Classification	- Numerical: 21 - Categorical: 4 - Derived: 4	Log Loss	- Precision - Recall - Accuracy - **F1 Score**

Results and Discussion

Algorithm	Train Results	Test Results
Majority Class Baseline	- Precision: 55.48% - Recall: 82.05% - Accuracy: 79.38% - **F1 Score: 70.35%**	- Precision: 64.98% - Recall: 80.61% - Accuracy: 80.61% - **F1 Score: 71.96%**
Random Class Baseline	- Precision: 67.13% - Recall: 49.87% - Accuracy: 50.05% - **F1 Score: 54.80%**	- Precision: 68.75% - Recall: 50.04% - Accuracy: 50.05% - **F1 Score: 55.14%**
Logistic Classification	- Precision: 77.68% - Recall: 80.63% - Accuracy: 80.63% - **F1 Score: 76.58%**	- Precision: 78.62% - Recall: 81.76% - Accuracy: 81.76% - **F1 Score: 78.15%**
Random Forest Classification	- Precision:76.14% - Recall: 78.36% - Accuracy: 78.34% - **F1 Score: 70%**	- Precision: 80% - Recall: 80.79% - Accuracy: 80.79% - **F1 Score: 72.54%**

Conclusion

- Best Performing Model: **Logistic Classification Model**
- Number of Features:
  - Numerical: 21
  - Categorical: 4
  - Derived: 4
- Top 10 Features:
  - Airline Performance
  - Distance
  - CRS Elapsed Time
  - CRS Departure Time
  - Hourly Wet Bulb Temperature
  - Hourly Relative Humidity
  - CRS Arrival Time
  - Hourly Visibility
  - Aircraft Flight Count
  - Hourly Wind Speed

Next Steps

- Apply dimensionality reduction techniques (PCA, regularization)
- Develop and test neural network classification model
- Fine tune the model

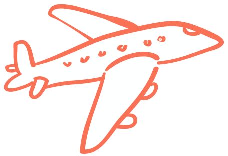
Thank you!

Why Investigate Flight Delay?

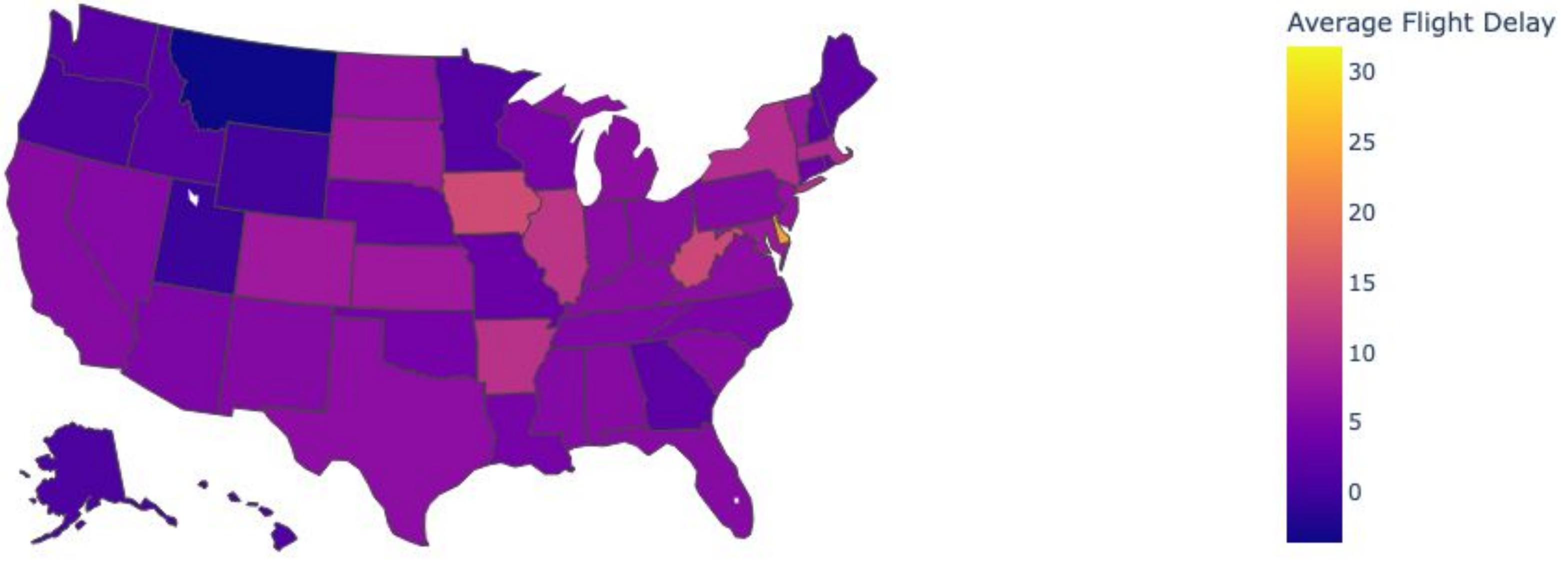
1. Improve Passenger Experience
2. Improve Airline Operational Efficiency
3. Identify Infrastructure Improvements
4. Advance Aviation Research & Development

Our Objective

In this project, we aim to develop a **predictive model** that can produce a **delay estimate** for flights within the United States given information about the flight, airport, and weather conditions.



EDA: Summary Visualizations



EDA: Missing Value Analysis

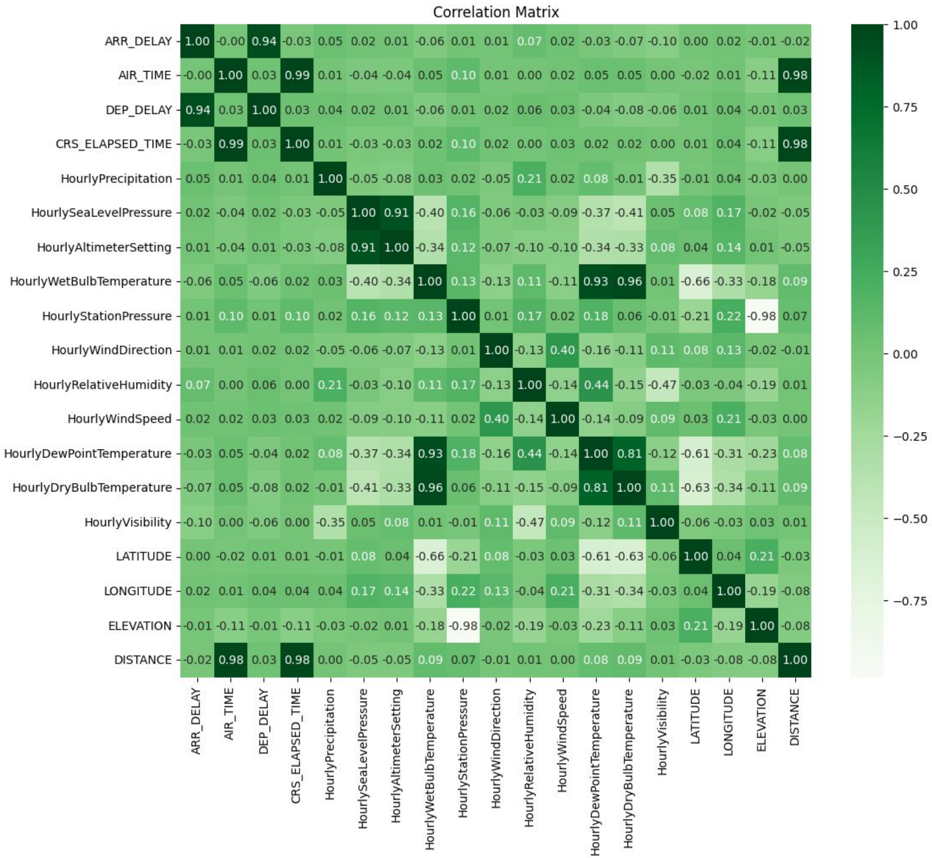
Column Name	count of missing value	Missing Value %
AWND	5811854	100
ShortDurationPrecipitationValue030	5811854	100
MonthlyMaxSeaLevelPressureValueTime	5811854	100
MonthlyDewpointTemperature	5811854	100
MonthlyGreatestPrecip	5811854	100
MonthlyGreatestPrecipDate	5811854	100
MonthlyGreatestSnowDepth	5811854	100
BackupName	3468275	59.67587968

- All the columns are removed above missing value percentage of 59.7% (BackupName)
- Read through glossary of features to confirm that removed features are not a critical part of our underlying project goals

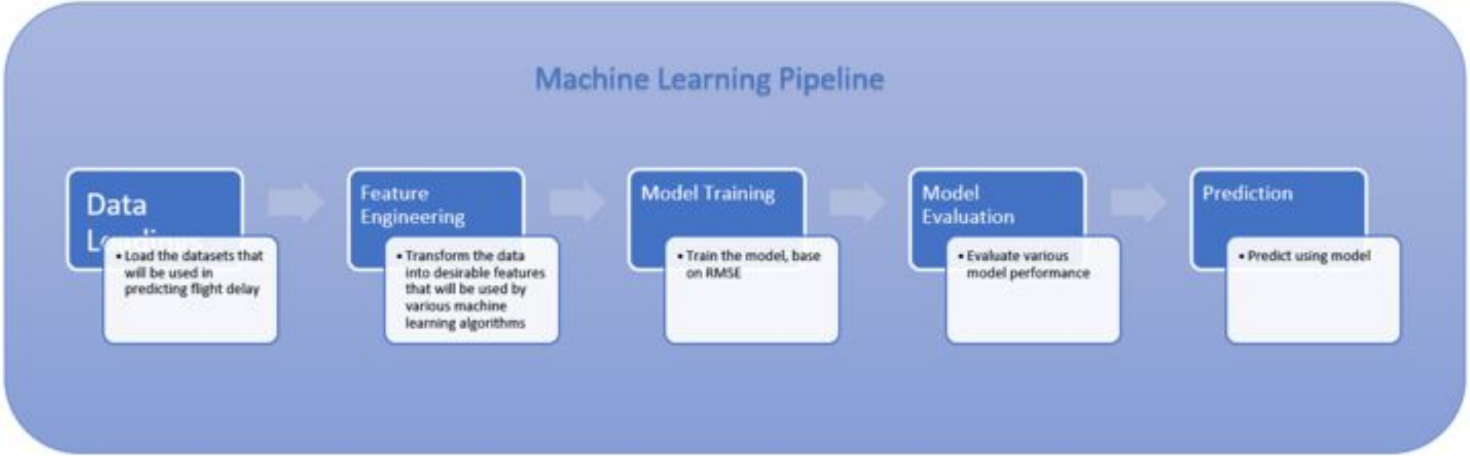
EDA: Missing Value Analysis

Feature Name (Some Missing %)	Action	Reasoning
CARRIER_DELAY, LATE_AIRCRAFT_DELAY, WEATHER_DELAY, SECURITY_DELAY, NAS_DELAY	Drop	High missing values; Data leakage
ARR_DELAY_GROUP, ARR_DEL15, ACTUAL_ELAPSED_TIME, AIR_TIME, ARR_DELAY_NEW, ARR_TIME, TAXI_IN, WHEELS_ON, WHEELS_OFF, TAXI_OUT, DEP_DELAY_GROUP, DEP_DELAY, DEP_DELAY_NEW, DEP_TIME, DEP_DEL15	Drop	Data Leakage
BackupElevation, BackupLatitude, BackupLongitude, BackupEquipment, BackupDistanceUnit, BackupDirection, BackupElements, BackupName	Drop	High missing values; less relevant.
BackupDistance	Impute 'Unknown'	High missing values; some distribution.
HourlyPressureChange, HourlyPressureTendency	Impute mean	Weather-related; might have predictive power.
HourlyPrecipitation, HourlySeaLevelPressure, HourlyAltimeterSetting	Impute mean	Weather-related; might have predictive power.
HourlyWetBulbTemperature, HourlyStationPressure, HourlyWindDirection, HourlyRelativeHumidity, HourlyWindSpeed, HourlyDewPointTemperature, HourlyDryBulbTemperature, HourlyVisibility	Impute mean	Weather-related; might have predictive power.
HourlySkyConditions, WindEquipmentChangeDate	Impute 'Unknown'	Weather-related; feature-engineering potential
TAIL_NUM	Impute 'Unknown'	Identifier; impute with placeholder.
REM	Impute 'Unknown'	Rare feature; impute with placeholder.
CRS_ELAPSED_TIME	Impute previous values; impute mean	Related to scheduled elapsed time; important for analysis.

EDA: Correlation Matrix



Modeling Pipeline



Baseline Model

X Variable	Features we derived
Y Variable	Arrival delay (Capped)
Model	OLS
Loss Function	MSE
R^2 Score	0.05

Conclusion & Next Steps

- Finalize feature engineering
- Dimensionality reduction (PCA, Lasso regularization)
- Build and test additional models
  - Classification model (predict bins of flight delay)
  - Neural network (model more non-linear relationships)