# AeroProphet

# Team 2.2


Hailee Schuele

hschuele@berkeley.edu

Landon Yurica

lyurica@berkeley.edu

Sreeram Ravinoothala

sreeram@berkeley.edu

Nick Johnson

nickjohnson@berkeley.edu

# Business Case

**Objective:** Predict departure delays greater than 15 minutes, 2 hours before takeoff

**Benefits:**

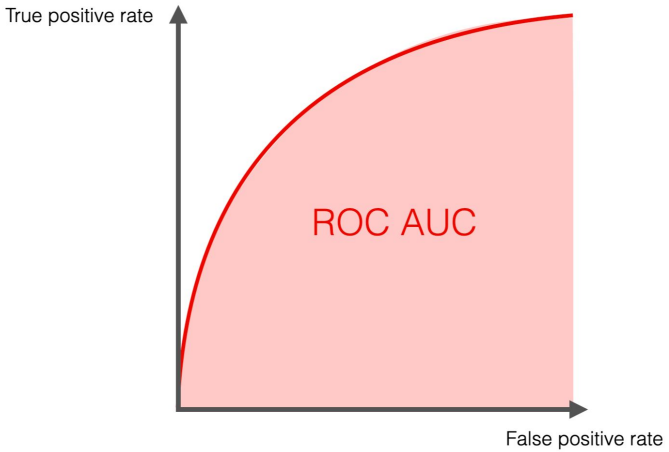
- For Airlines :
  - Proactively address issues
  - Improve efficiency
- For Passengers:
  - Adapt plans
  - Make informed decisions

# Metric Selection

**Primary Metric:**

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

**Secondary Metric:**



# Final Result

**Out of Sample Performance:**

Primary Metric:

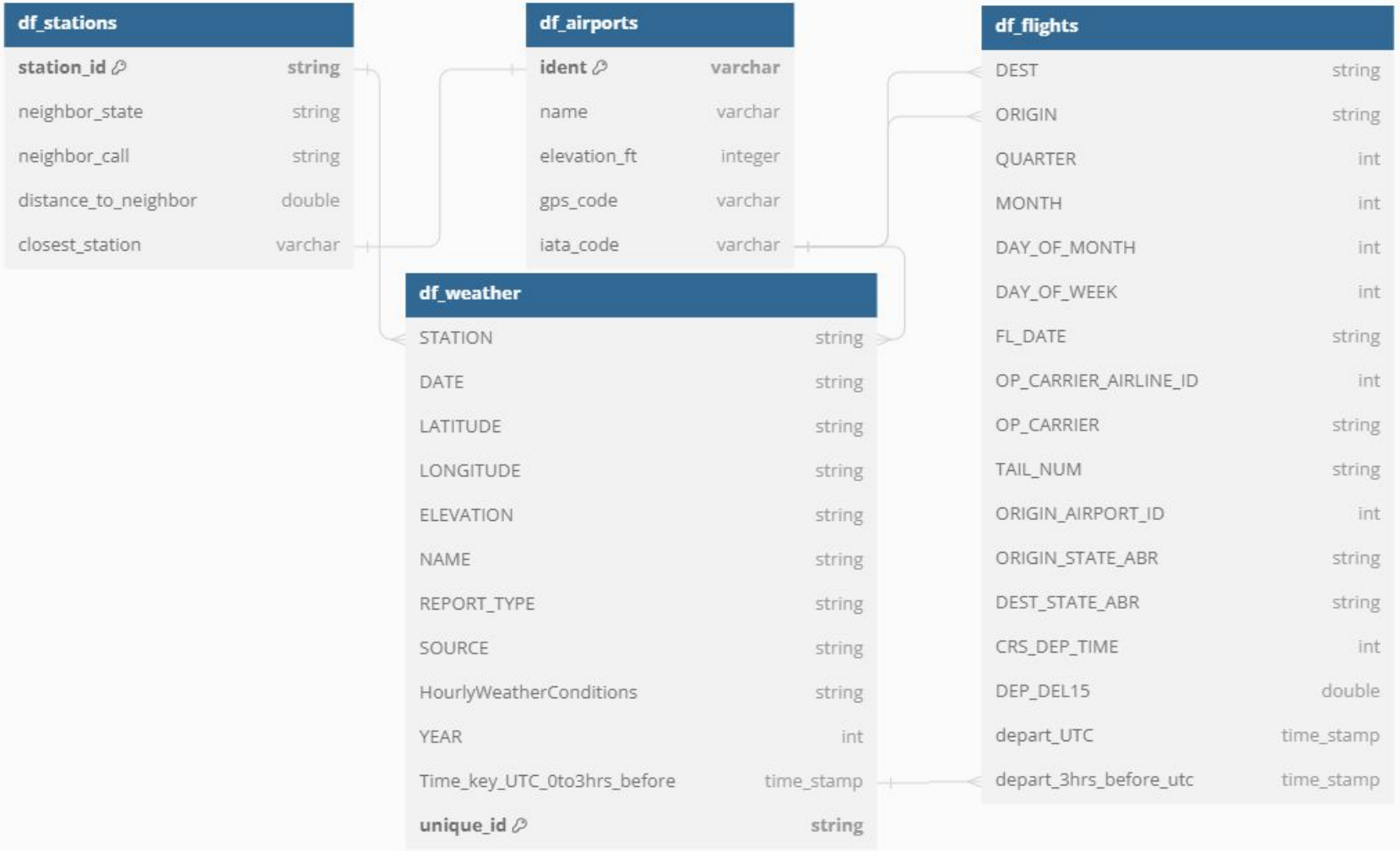
F1 Score = 0.803

Secondary Metric:

AUC = 0.734

# Datasets

- 2015-2021US Flight records
- 2015-2021 International Weather records
- Weather Station location data
- Airport location data


df_stations
station_id string
neighbor_state string
neighbor_call string
distance_to_neighbor double
closest_station varchar
df_airports
ident varchar
name varchar
elevation_ft integer
gps_code varchar
iata_code varchar
df_flights
DEST string
ORIGIN string
QUARTER int
MONTH int
DAY_OF_MONTH int
DAY_OF_WEEK int
FL_DATE string
OP_CARRIER_AIRLINE_ID int
OP_CARRIER string
TAIL_NUM string
ORIGIN_AIRPORT_ID int
ORIGIN_STATE_ABR string
DEST_STATE_ABR string
CRS_DEP_TIME int
DEP_DEL15 double
depart_UTC time_stamp
depart_3hrs_before_utc time_stamp
df_weather
STATION string
DATE string
LATITUDE string
LONGITUDE string
ELEVATION string
NAME string
REPORT_TYPE string
SOURCE string
HourlyWeatherConditions string
YEAR int
Time_key_UTC_0to3hrs_before time_stamp
unique_id string

# Data Join Details

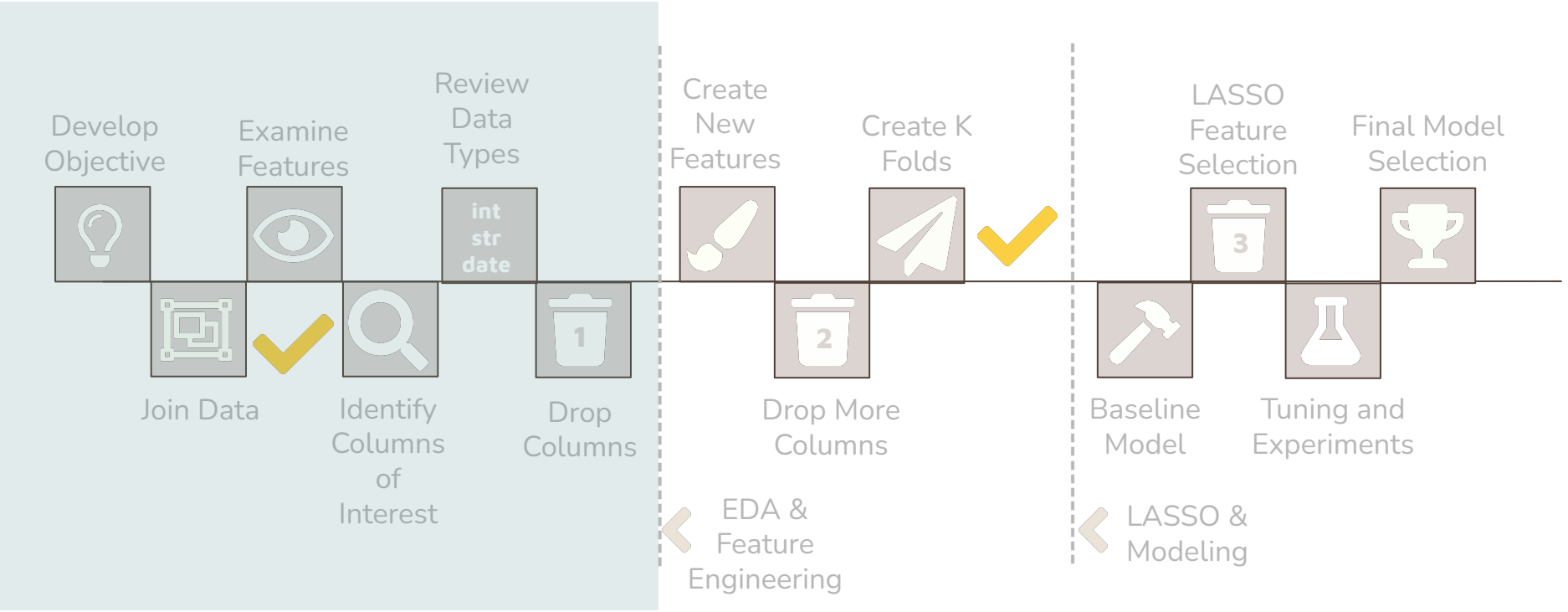
| Table | Rows | Columns | Memory (GB) | Run Time (HH:MM:SS) |
|---|---|---|---|---|
| df_flights | 74,177,433 | 109 | 2.93 | None |
| df_weather | 898,983,399 | 124 | 35.05 | None |
| df_stations | 5,004,169 | 12 | 1.3 | None |
| df_airports | 57,421 | 12 | 0.01 | None |
| df_FSW | 72,515,921 | 292 | 43.5 | 02:13:00 |
| df_FSW_Clean | 64,457,088 | 87 | 64.78 | 06:18:00 |

Cluster Details: DBR 13.3 LTS ML, Spark 3.4.1, Scala 2.12, Standaard_DS3_v2, 14GB, 4 Cores, Standard_DS3_v2, 84GB, 24 Cores, 1-6 workers

# Data Cleaning

- Dropped columns containing:
  - Substantial missing data
  - Potential data leakage
  - Redundant information
- Dropped duplicate rows
- Mean imputation

# Data Join & Cleaning



# Feature Engineering

**Existing Engineered Features**

**previous_flight_delay**
- Delay time of the previous flight

**state_to_region**
- Convert state abbreviations to geographic region

**Plane_Delays_last_24h**
- Sum of departure delays for each plane (based on tail number) within the last 24 hours

**airport_delays_in_previous_24_hours**
- The number of flights delayed at the airport in the last 24 hours

**days_to_nearest_holiday**
- The number of days to the nearest holiday

**Newly Engineered or Added**

➔ Added min & max hourly weather values

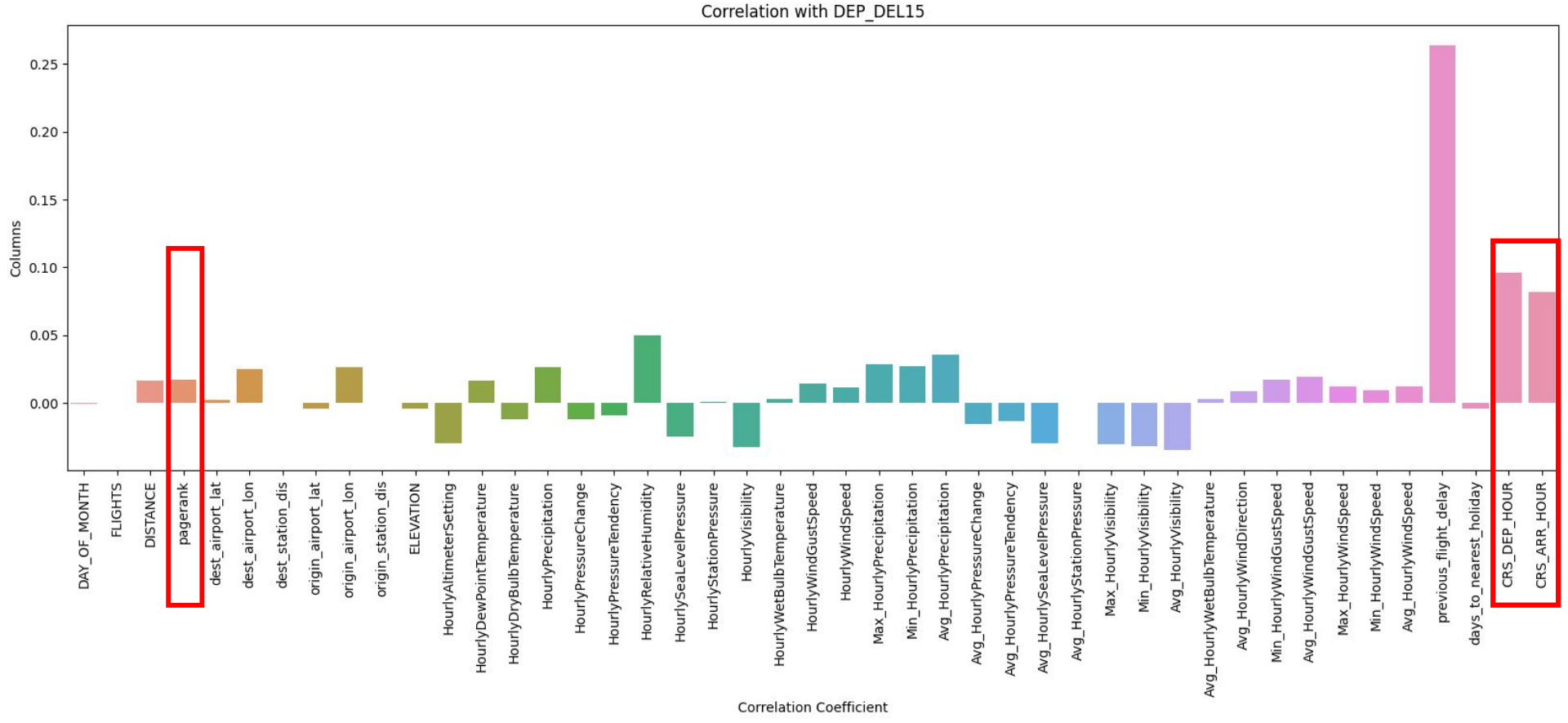
➔ Airport Page Rank

➔ Airport Size

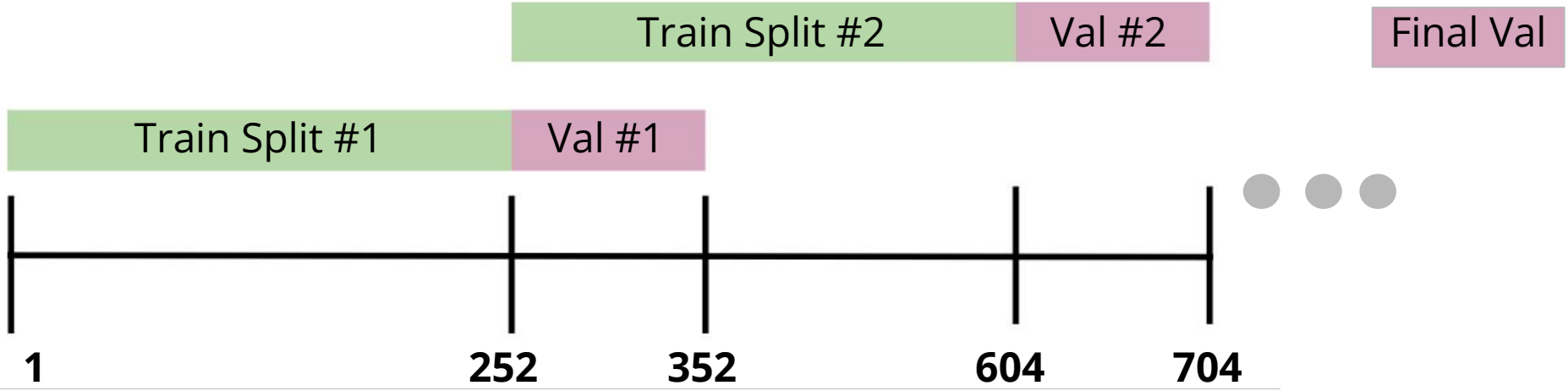
➔ Scheduled depart and arrival hour (local time)

➔ Class + Recency

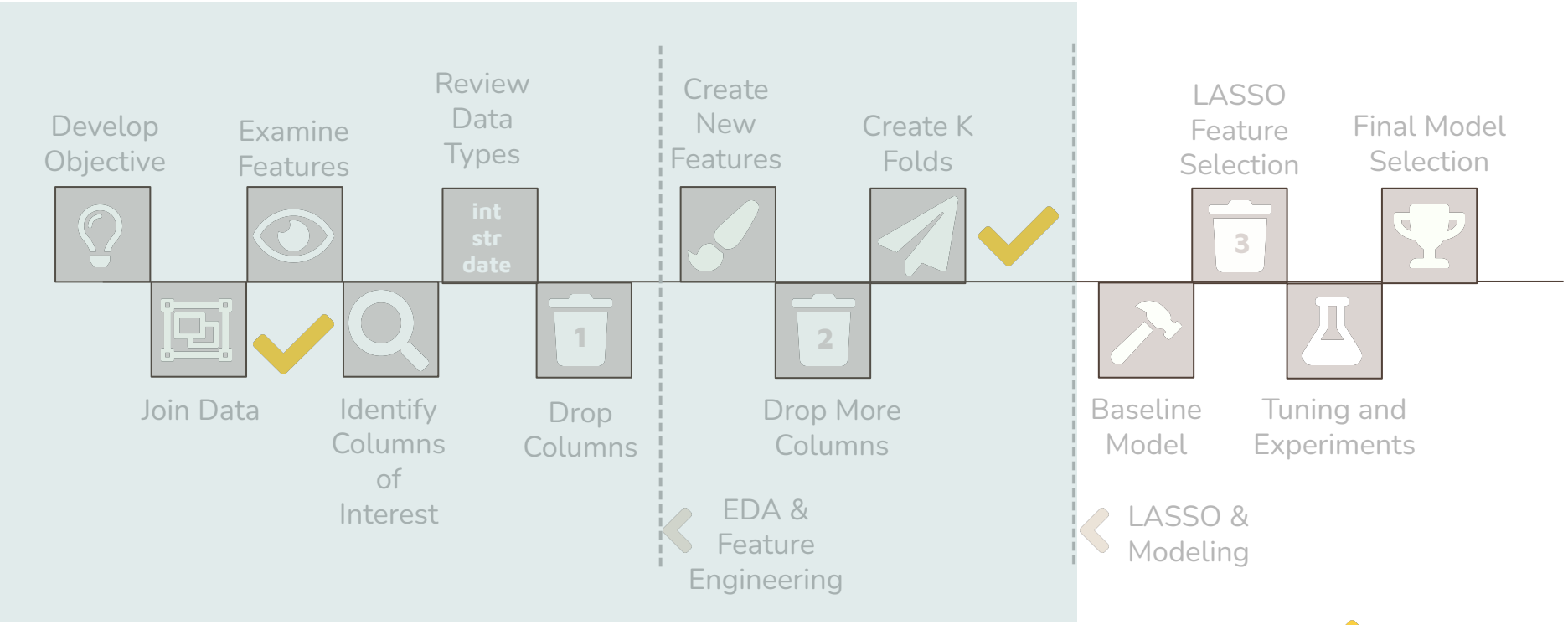
# EDA


Correlation with DEP_DEL15
Columns
0.25
0.20
0.15
0.10
0.05
0.00
DAY_OF_MONTH
FLIGHTS
DISTANCE
pagerank
dest_airport_lat
dest_airport_lon
dest_station_dis
origin_airport_lat
origin_airport_lon
origin_station_dis
ELEVATION
HourlyAltimeterSetting
HourlyDewPointTemperature
HourlyDryBulbTemperature
HourlyPrecipitation
HourlyPressureChange
HourlyPressureTendency
HourlyRelativeHumidity
HourlySeaLevelPressure
HourlyStationPressure
HourlyVisibility
HourlyWetBulbTemperature
HourlyWindGustSpeed
HourlyWindSpeed
Max_HourlyPrecipitation
Min_HourlyPrecipitation
Avg_HourlyPrecipitation
Avg_HourlyPressureChange
Avg_HourlyPressureTendency
Avg_HourlySeaLevelPressure
Avg_HourlyStationPressure
Max_HourlyVisibility
Min_HourlyVisibility
Avg_HourlyVisibility
Avg_HourlyWetBulbTemperature
Avg_HourlyWindDirection
Min_HourlyWindGustSpeed
Avg_HourlyWindGustSpeed
Max_HourlyWindSpeed
Min_HourlyWindSpeed
Avg_HourlyWindSpeed
previous_flight_delay
days_to_nearest_holiday
CRS_DEP_HOUR
CRS_ARR_HOUR
Correlation Coefficient

# Cross Folds


Train Split #2
Val #2
Final Val
Train Split #1
Val #1
1
252
352
604
704

# EDA & Feature Engineering



# Modeling Workflow


- 70 features
- Includes derived features on folds
- Validation data included
- Used Vector representation of features

- 64 features
- Excluded derived
- On folds
- Retained features kept in at least one fold
- Got 38 features

- 38 Lasso features
- Additionally 5 Derived features
- Ran on both folds as well as final validation set

- One fold using lasso dataset
- Included derived columns
- No significant impact with Hyper params
- Used multiple seeds, avg for eval
- Ran on all folds
- Numtree=10, maxdepth=5

- One fold using lasso dataset
- Included derived columns
- No significant impact with Hyper params
- Used multiple seeds, avg for eval
- Ran on all folds
- iter=10, stepsize=1.0

# Neural Network Experiments

## MLP

- Experimented with different seeds, iterations, number of nodes in hidden layers with a 10% dataset
- Results showed maxiter=100 provided good result
- Both single hidden layer and two hidden layers come back with close results
- Lasso data used

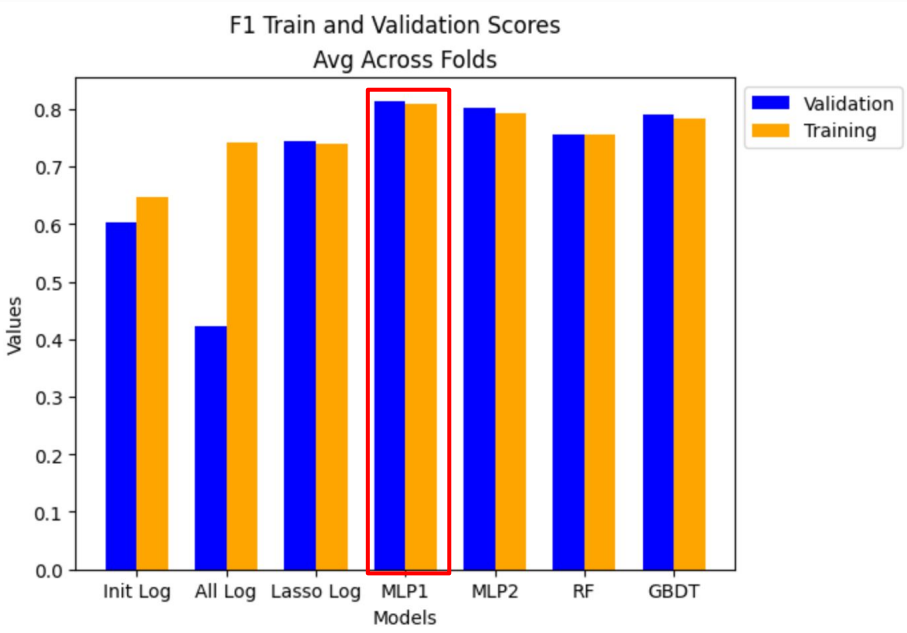
## LSTM

- Ran for 20 epochs, on all folds. Loss converged at 10 epochs
- Used BCELoss, Adam optimizer
- Captured scores on fold validations and final validation set
- Used only 10% of the data
- Taking too long
- Future

## Final MLP

- Single hidden layer

# Intermediary Model Results




| | Train F1 | Validate F1 |
|---|---|---|
| Initial Logistic | 0.648 | 0.604 |
| Secondary Logistic | 0.741 | 0.422 |
| LASSO Logistic | 0.740 | 0.743 |
| MLP 2 Hidden Layers | 0.792 | 0.801 |
| Random Forest | 0.755 | 0.756 |
| Gradient Boosted Decision Trees | 0.783 | 0.791 |
| MLP 1 Hidden Layer | 0.808 | 0.814 |

# Final Model Results

| | Train F1 | Test F1 |
|---|---|---|
| **Final MLP** | 0.805 | 0.803 |

# Conclusion

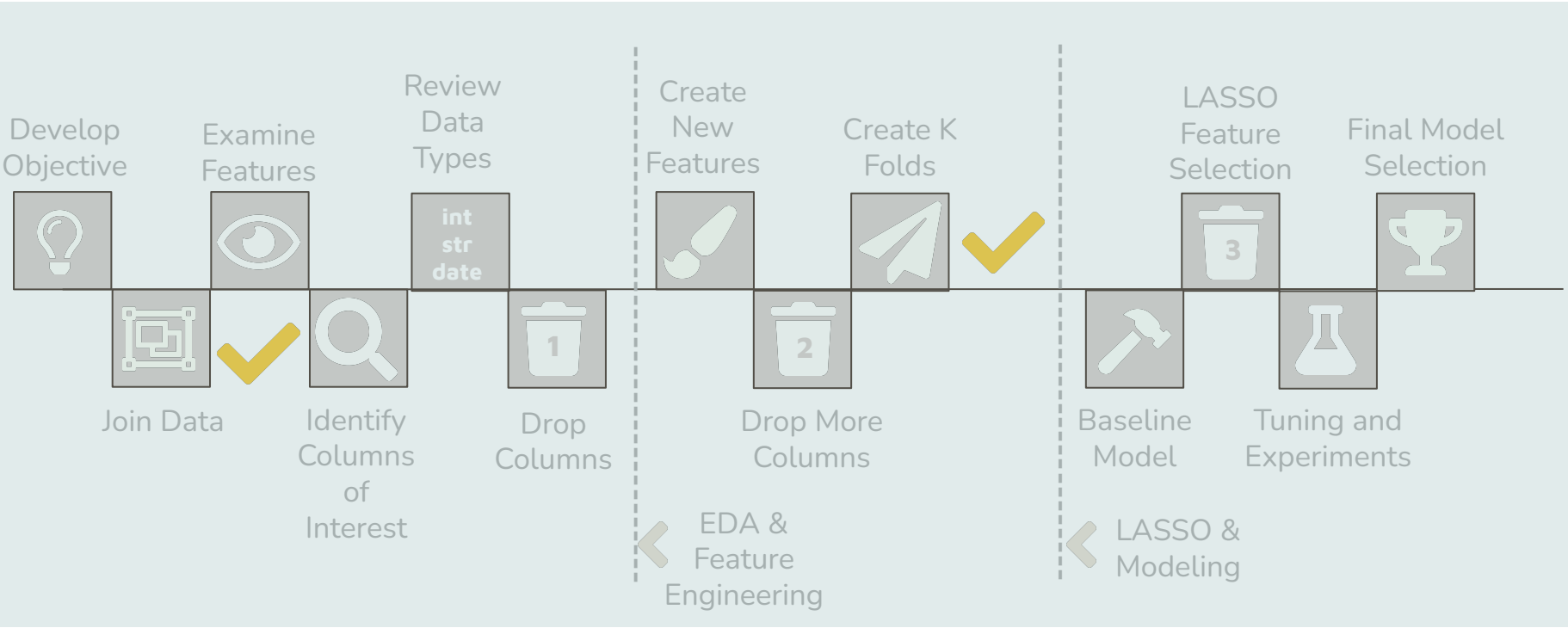
**Final Model**

- 38 LASSO Features + 5 Derived Features
- 35% increase in F1 score from first to last model

**Next Steps**

➔ More Derived Features
➔ Additional Hyperparameter Tuning
➔ Ensemble Methods
➔ Deep Learning and More Complex Models

# Thank You!


Develop Objective
Examine Features
Review Data Types
int
str
date
Create New Features
Create K Folds
LASSO Feature Selection
3
Final Model Selection
Join Data
Identify Columns of Interest
1
Drop Columns
2
Drop More Columns
Baseline Model
Tuning and Experiments
EDA & Feature Engineering
LASSO & Modeling
Data Checkpoint