Predicting Flight Delays

W261 Spring 2024
Machine Learning at Scale
Group 4 | Team 3_4

Meet The Team

Daphne Lin

Nicole Liu

Tanmay Mahapatra

Jun Park

Agenda

- Project Problem
- EDA
- Feature Engineering
- Modeling Pipeline
- Models and Experiment
  Results
- Conclusion and Next Steps

Project Statement and Description

Issue

- Impact on Schedules/Routes
  - Delays, Diversion, Cancellation
- Impact on Cargo
  - Delayed Delivery
  - Damage to perishable items
- Impact on Airports
  - Congestion and capacity
- Impact on Passengers
  - Accommodation costs

Proposed Project

- **Goal:** Predict delays in advance to drive actionable insights for airlines and airports
- **Method:** Develop predictive models to forecast departure delays with logistic regression as a baseline
- **Main Metrics:** F-2 = 2, Recall (Reduce False Negatives)
- **Target audience:** Airport authority, Airline Carriers, Passengers

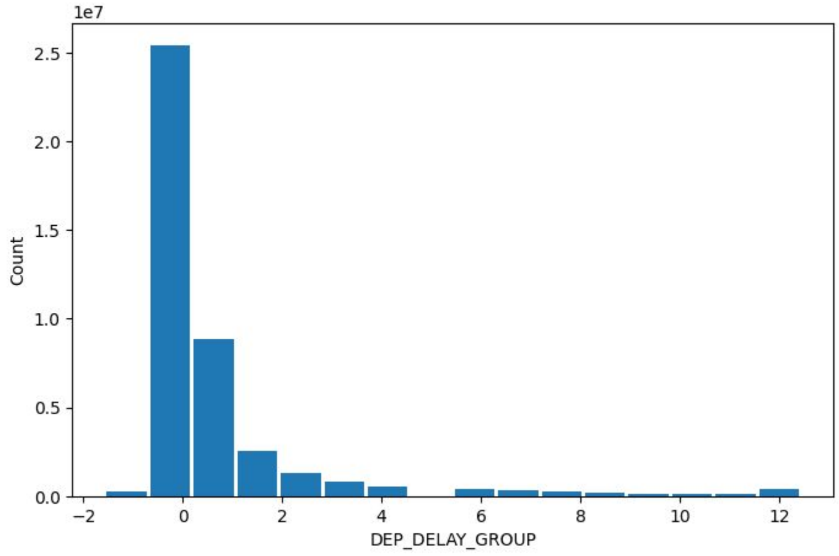
EDA - Flights

- Raw airlines dataset contains 74,177,433 records
- After removing duplicates, the dataset consists of 42,430,592 records
- Approximately 2.02% of flights are cancelled flights
- Cancelled flights, represented by null values in the DEP_DEL15 feature, are removed from the dataset due to irrelevance and minimal occurrence

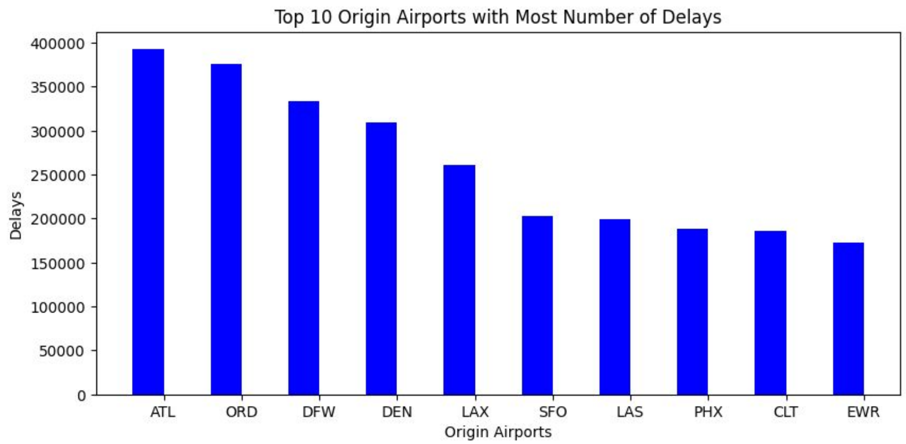
	Metric	Value
0	Total Flights	42430592
1	Number of Unique Carriers	20
2	Number of Unique Airports	388
3	Flights delayed over 15mins	7119338
4	Delayed Flights %	16.78%
5	On-Time Flights %	81.20%
6	Canceled Flights %	2.02%

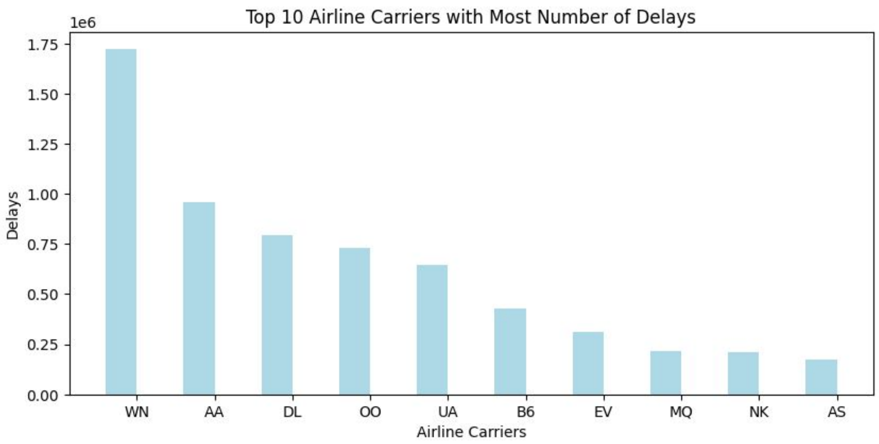
EDA - Flights Contd.

- Class imbalance: 16.78% of flights are delayed by more than 15 minutes, while 81.20% are on time, indicating potential prediction bias towards on-time flights
- Future steps involve addressing class imbalance by applying a balancing ratio using a weight column
- The DEP_DELAY_GROUP field displays delay times grouped in 15-minute intervals, skewed towards shorter delays



EDA - Origin Airports and Carriers

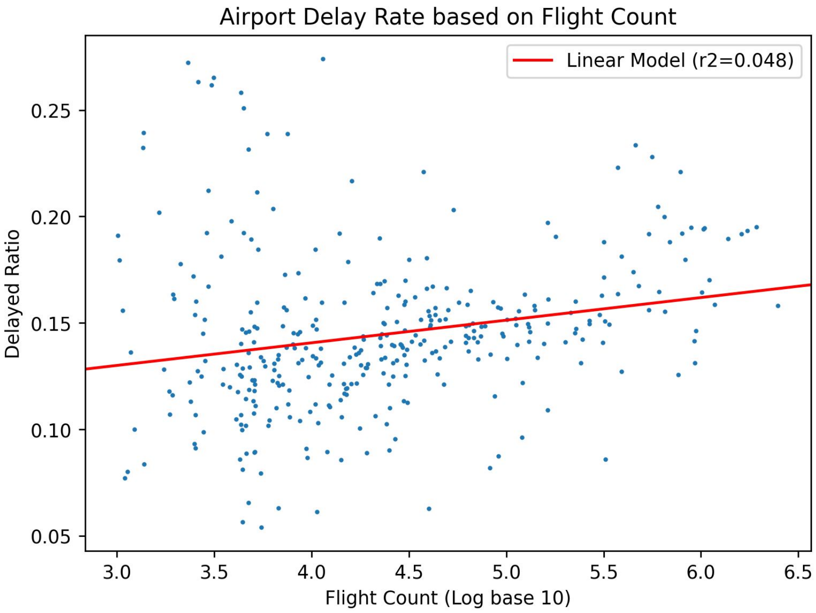




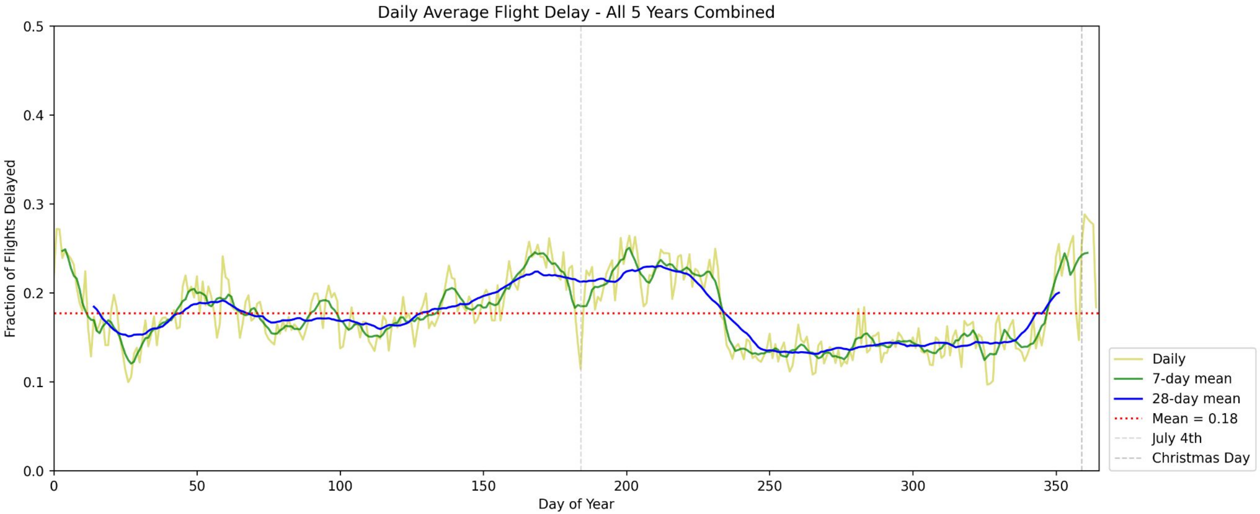
- 5 airports with the highest number of delays: Atlanta (ATL), Orlando (ORD), Dallas-Fort Worth (DFW), Denver (DEN), Los Angeles (LAX)
- Top 5 carriers (IATA Code): Southwest Airlines (WN), American Airlines (AA), Delta Airlines (DL), SkyWest Airlines (OO), United Airlines (UA)

EDA - Origin Airports Contd.

- Scatter plot analysis reveals a subtle, increasing linear relationship between airport business and the proportion of delayed flights
- Significant variance within the data, suggesting that the origin airport as a feature in predictive models could be helpful
- Feature engineering involves integrating average delay by airport before departure into the predictive modeling process
- Understanding the relationship between airport business and flight delays aids in optimizing feature selection for improved delay predictions



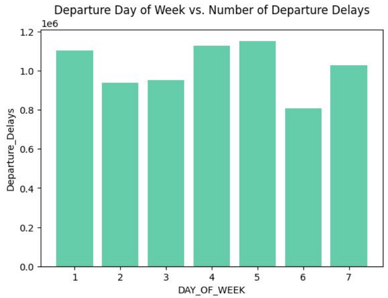
EDA - Holiday Impact

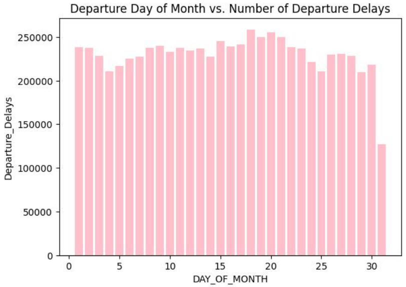


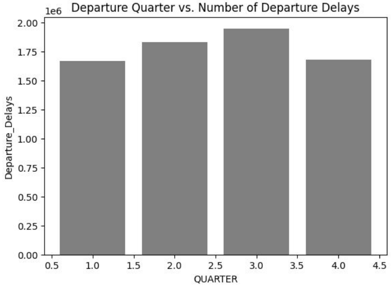
- 2 dips denoted by dotted lines on 2 occasions (holidays): 04/07, 12/25
- Indicators highlight that travel is usually before or after a holiday

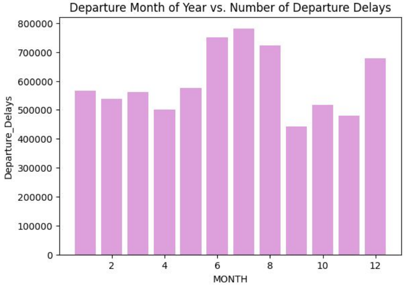
EDA - Time factors

- Thursday and Friday have slightly more delays in a week.
- Day of Month has close to uniform distribution.
- Slightly more delays in Q2 and Q3
- June to August and December have more delays in a year









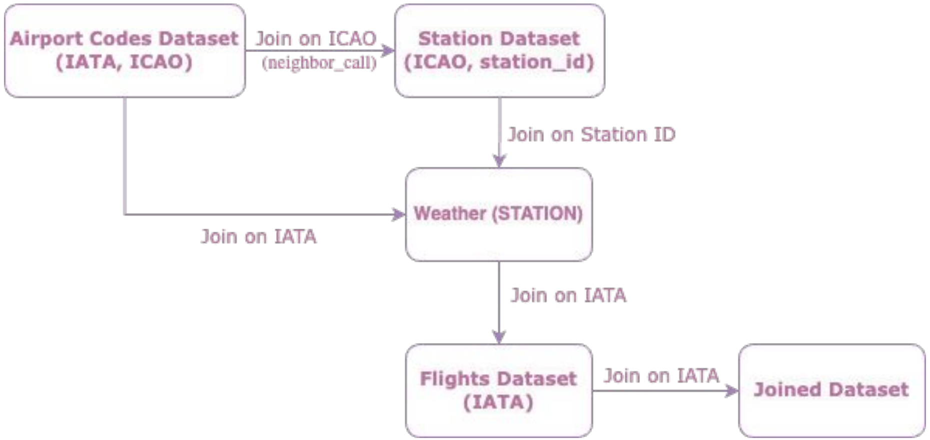
EDA - Weather

- Weather dataset contains 898,983,399 records
- There were no duplicates
- 15,027 unique weather stations
- Earliest recorded date within the considered dataset 01/01/2015
- Earliest recorded date within the considered dataset 12/31/2021
- Most of the data is null

	Metric	Value
0	Total Weather Records	898983399
1	Number of Unique Weather Stations	15027
2	Earliest Date of recorded weather	2015-01-01
3	Latest Date of recorded weather	2021-12-31

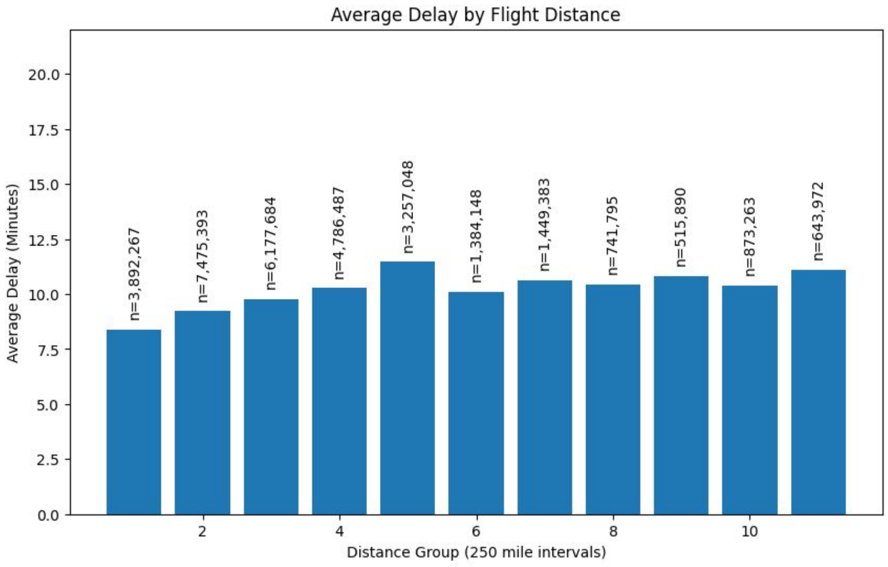
Joined Dataset

- Airport Codes dataset shared with us that houses IATA as well as ICAO codes
- Airport codes dataset can be joined with Station dataset by joining on ICAO code to 'neighbor_call'
- Station dataset can be joined with Weather by joining 'station_id' to STATION
- Airport Codes dataset can be combined with the Weather dataset based on IATA code
- Join the result with Flights dataset on IATA code to create final Joined Dataset



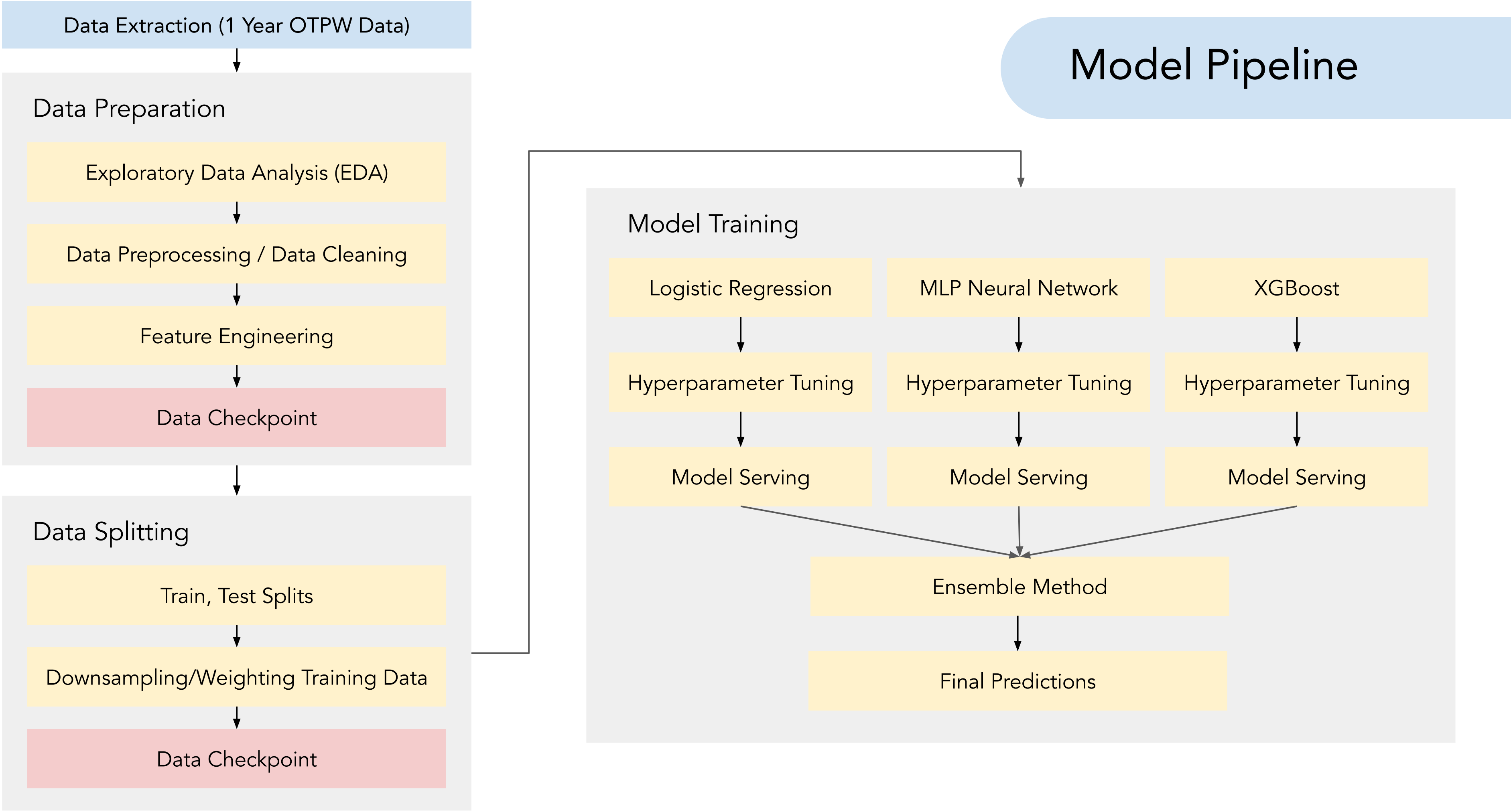
EDA - Joined Dataset

- Histogram of Average Delay by Flight Distance showcases the relationship between departure delay and flight distance, organized into 250-mile intervals
- Distribution demonstrates a **nearly uniform pattern across flight distances**, with group 2 accounting for the largest number of flights (n = 7,475,393)
- Notably, flights categorized as **middle-distance (group 5) exhibit a slightly higher average delay**, with an average delay time of about 11 minutes



New Feature Creation

Is near a major holiday	• Recorded dates of major holidays identified in the EDA (Christmas, Thanksgiving, New Years, 4th of July) • Days within a 3 day window of the holiday were recorded as 'is_near_holiday'
% delays at origin airport (2-4 hours before departure)	• Calculated the % of delays at the origin airport for flights within the 2-4 hour window before departure.
% delays at dest airport (2-4 hours before departure)	• Calculated the % of delays at the destination airport for flights within the 2-4 hour window before departure.
Average tail delay time (2 hours before departure, previous 4 flights)	• Calculated average tail number delay time for the last 4 flights 2 hours prior to departure.


Model Pipeline
Data Extraction (1 Year OTPW Data)
Data Preparation
Exploratory Data Analysis (EDA)
Data Preprocessing / Data Cleaning
Feature Engineering
Data Checkpoint
Data Splitting
Train, Test Splits
Downsampling/Weighting Training Data
Data Checkpoint
Model Training
Logistic Regression
MLP Neural Network
XGBoost
Hyperparameter Tuning
Hyperparameter Tuning
Hyperparameter Tuning
Model Serving
Model Serving
Model Serving
Ensemble Method
Final Predictions

Feature Transformation

Dataset: OTPW 2015 full year

Stage 1: Feature elimination

- Co-linearity with other features
- Info not known 2 hrs in advance
- Irrelevant
- Contains a lot of null values

Stage 2: Data processing

- Handle null values
- Transform categorical to one-hot encoding
- Extract hour data

Stage 3: Data split

Train / Test split based on time series

Stage 4: Model training

- 18-22 base features
- 1 binary prediction

Stage 5: Model iteration and improvement

- Metrics (focus on recall, f-2 score)
- Perform experiments
- Hyperparameter (Grid search)

Experiment Type	Data Source	Input Features	Train Results	Test Results
Baseline: no class weights	1 Year OPTW (no downsampling)	**10 Numerical** (Flight Distance, Hourly Weather) **10 Categorical** (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.0335 F-2: 0.0409 F-1: 0.0614
With class weights, regParam = 0.1, elasticNetParam = 0.5, maxIter = 20	1 Year OPTW (no downsampling)	**10 Numerical** (Flight Distance, Hourly Weather) **10 Categorical** (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.6636 F-2: 0.4775 F-1: 0.3361
With class weights; regParam = 0.1, elasticNetParam = 0.5, maxIter = 20, add tail number, flight number as categorical features	1 Year OPTW (no downsampling)	**10 Numerical** (Flight Distance, Hourly Weather) **12 Categorical** (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.6637 F-2: 0.4775 F-1: 0.3361
(no class weights after downsampling) regParam = 0.1, elasticNetParam = 0.5, maxIter = 20	5 Year OPTW (with downsampling)	**4 Feature Engineered** (Holiday, Delay at Origin/Destination, Avg Delay for Plane) + **10 Numerical** (Flight Distance, Hourly Weather) **11 Categorical** (Time, Airline Info, Origin, Destination)	Recall: 0.5473 F-2: 0.5684 F-1: 0.6032	Recall: 0.5593 F-2: 0.4798 F-1: 0.3954
GridSearch; regParam = 0.01, elasticNetParam = 0.0, maxIter = 10	5 Year OPTW (with downsampling)	**4 Feature Engineered** (Holiday, Delay at Origin/Destination, Avg Delay for Plane) + **10 Numerical** (Flight Distance, Hourly Weather) **11 Categorical** (Time, Airline Info, Origin, Destination)	Recall: 0.5902 F-2: 0.6054 F-1: 0.6297	Recall: 0.6026 F-2: 0.6054 F-1: 0.4070

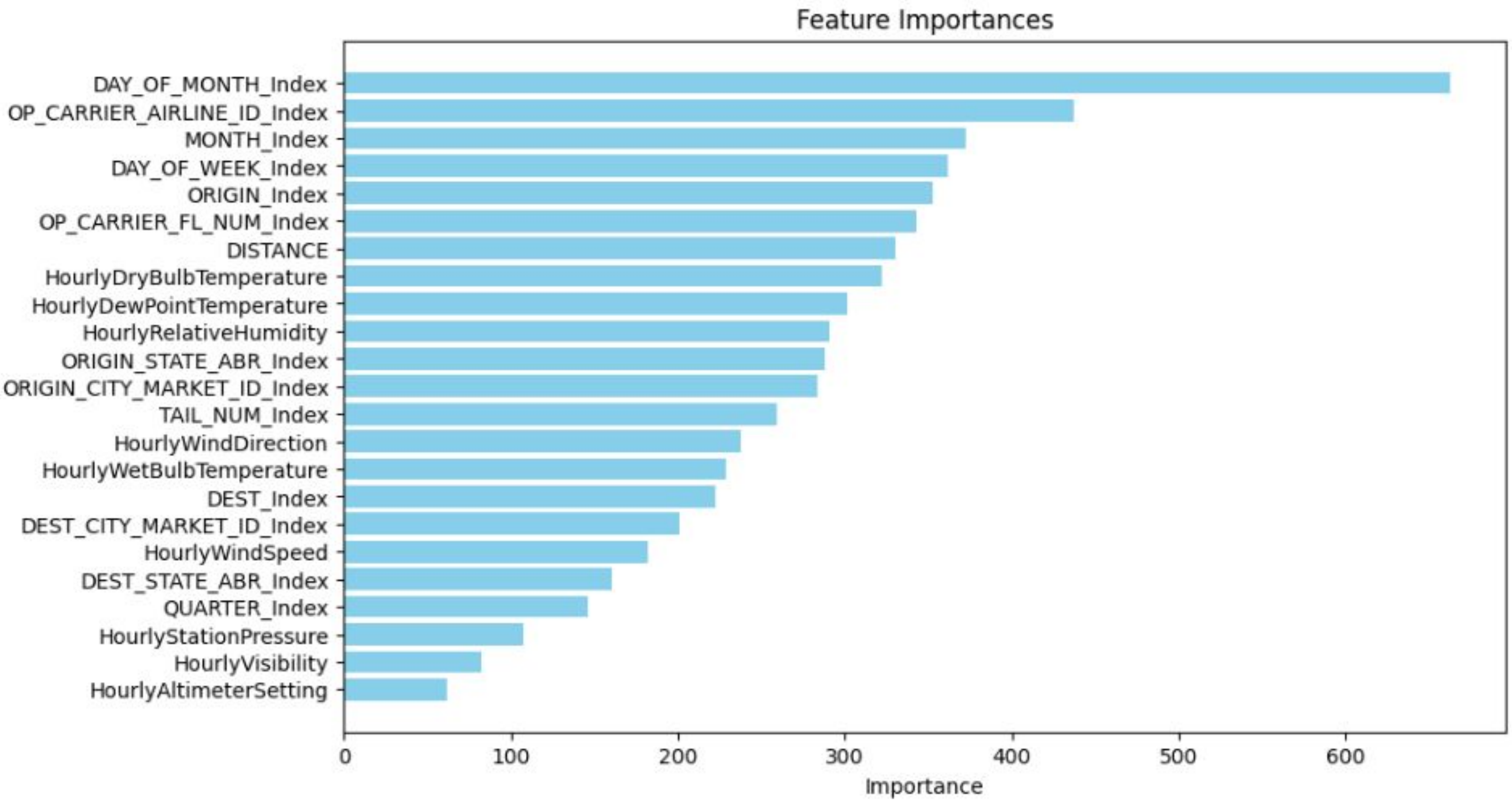
Experiment	Input Features	Train Results	Test Results
MLP with 1 Hidden Layer (size 4) maxIter = 50, stepSize = 0.03, blockSize = 128 (1 hour training time) Neural Network Architecture: MLP-718 - 4 Sigmoid - 2 Softmax	**10 Numerical** (Flight Distance, Hourly Weather) **8 Categorical** (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.5577 F-2: 0.4031 F-1: 0.2847
MLP with 2 Hidden Layers (both size 4) maxIter = 50, stepSize = 0.03, blockSize = 128 (1 hour training time) Neural Network Architecture: MLP-718 - 4 Sigmoid - 4 Sigmoid - 2 Softmax	**10 Numerical** (Flight Distance, Hourly Weather) **8 Categorical** (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.4924 F-2: 0.3793 F-1: 0.2821
GridSearch CV MLP w/ 1 Hidden Layer maxIter = 50, stepSize = 0.01, blockSize = 128 (4 hour training time) Neural Network Architecture: MLP-718 - 10 Sigmoid - 2 Softmax	**10 Numerical** (Flight Distance, Hourly Weather) **8 Categorical** (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.6230 F-2: 0.4237 F-1: 0.2863

Experiment	Input Features	Train Results	Test Results
MLP with 1 Hidden Layer (size 4) maxIter = 50, stepSize = 0.03, blockSize = 128 (1 hour training time) Neural Network Architecture: MLP-804 - 4 Sigmoid - 2 Softmax	**3 Feature Engineered** (Holiday, Delay at Origin/Destination) **10 Numerical** (Flight Distance, Hourly Weather) **8 Categorical** (Time, Airline Info, Origin, Destination)	Recall: 0.9999 F-2: 0.8358 F-1: 0.6707	Recall: 1.000 F-2: 0.5378 F-1: 0.3176
MLP with 2 Hidden Layers (both size 4) maxIter = 50, stepSize = 0.03, blockSize = 128 (1 hour training time) Neural Network Architecture: MLP-804 - 4 Sigmoid - 4 Sigmoid - 2 Softmax	**3 Feature Engineered** (Holiday, Delay at Origin/Destination) **10 Numerical** (Flight Distance, Hourly Weather) **8 Categorical** (Time, Airline Info, Origin, Destination)	Recall: 1.0000 F-2: 0.8358 F-1: 0.6707	Recall: 1.0000 F-2: 0.5378 F-1: 0.3176
GridSearch CV MLP w/ 1 Hidden Layer maxIter = 50, stepSize = 0.01, blockSize = 128 (3 hour training time) Neural Network Architecture: MLP-804 - 10 Sigmoid - 2 Softmax	**3 Feature Engineered** (Holiday, Delay at Origin/Destination) **10 Numerical** (Flight Distance, Hourly Weather) **8 Categorical** (Time, Airline Info, Origin, Destination)	Recall: 0.5508 F-2: 0.5712 F-1: 0.6049	Recall: 0.5599 F-2: 0.4807 F-1: 0.3967

Experiment Type	Data Source	Input Features	Train Results	Test Results
With class weights, Num_round = 50, max_depth=6	1 Year OPTW (no downsampling)	10 Numerical (Flight Distance, Hourly Weather) 10 Categorical (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.4588 F-2: 0.3786 F-1: 0.3000
With class weights, Num_round = 100, max_depth=6,	1 Year OPTW (no downsampling)	10 Numerical (Flight Distance, Hourly Weather) 10 Categorical (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.4588 F-2: 0.3786 F-1: 0.3000
With class weights, Num_round = 100, scalePosWeight=4, min_child_weight=1, max_depth=6, subsample=0.8, colsample_bytree=0.8	1 Year OPTW (no downsampling)	10 Numerical (Flight Distance, Hourly Weather) 10 Categorical (Time, Airline Info, Origin, Destination)	[not recorded]	Recall: 0.4588 F-2: 0.3786 F-1: 0.3000
(no class weights after downsampling) Num_round = 20, min_child_weight=1, max_depth=6, subsample=0.8, colsample_bytree=0.8	5 Year OPTW (with downsampling)	4 Feature Engineered (Holiday, Delay at Origin/Destination, Avg Delay for Plane) + 10 Numerical (Flight Distance, Hourly Weather) 11 Categorical (Time, Airline Info, Origin, Destination)	Recall: 0.6114 F-2: 0.6231 F-1: 0.6415	Recall: 0.6348 F-2: 5192 F-1: 0.4079
Early stopping: Same parameters as above + except num_round = 1000, num_early_stopping_rounds = 10, eval_metric = "logloss", maximize_evaluation_metrics = False	5 Year OPTW (with downsampling)	4 Feature Engineered (Holiday, Delay at Origin/Destination, Avg Delay for Plane) + 10 Numerical (Flight Distance, Hourly Weather) 11 Categorical (Time, Airline Info, Origin, Destination)	Train: Recall: 0.6259 F-2: 0.6340 F-1: 0.6466 Val: Recall: 0.6673 F-2: 0.6622 F-1: 0.6548	Recall: 0.6066 F-2: 0.4979 F-1: 0.3924

Experiment	Input Features	Test Results
Ensemble Method	GridSearch Logistic Regression Model GridSearch MLP Neural Network Model XGBoost w/o Early Stopping Model	Recall: 0.3229 F-2: 0.2853 F-1: 0.2428

Existing Feature Importance Scores



Conclusion / Next Steps

Conclusion

- Added 4 new features
- 3 sets of models
- Best models (logistic regression)
    - 10 numeric + 10 categorical
    - Recall: 0.8383; F-2: 0.5009
- Top features
    - Date related
    - Airline ID
    - Origin

Next Steps

- Include new features into model training
- Grid search on XGBoost
- Cross validation
- Clean up end-to-end pipeline
- Gap analysis

Appendix
Dataset

Datasets

- Airlines Data:
  - https://www.transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr
  - Airline performance data from the TranStats data compilation, specifically focusing on on-time performance of passenger flights (2015) supplied by the Department of Transportation (DOT)
- Weather Data:
  - https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516
  - Weather dataset (2015) gathered from the National Oceanic and Atmospheric Administration (NOAA) repository
- Stations Data:
  - Airline Station dataset with the necessary keys for merging Flight data with Weather data
- OTPW Data (Airlines + Weather):
  - Joined dataset provided to us that combines the Airlines and Weather datasets

THANK YOU

Questions?