Delay Dissatisfaction (D2) Project Updates

Team 1, Emerald Airlines, UI/UX Feature Research Dept.

04/17/2024

Meet our Team!

Nathan Arias
nathanarias@berkeley.edu

Thomas Dolan
tdolan@berkeley.edu

Maegan Kornexl
mkornexl@berkeley.edu

I-Hsiu Kao
ihsiukao@berkeley.edu

Presentation Outline

- Project Overview
  - Background and goals
  - Company/customer relevance
- Exploratory Data Analysis (EDA) and Feature Selection
- Model Development
- Results and Discussion
- Limitations and Future Directions

1.

The Project

How can we improve the overall passenger experience around delays?

7,000 to 9,000 flights are delayed daily, making up 25% of all flights annually.

Delays cost airlines ~$20,000/hour and passengers ~$47/hour, leading to significant financial costs and long-term damage to both customer loyalty and brand reputation. [1,2]

2023 GEM Survey Results

- Would you be more accepting of a 15+ minute delay if you were notified beforehand that it was likely to happen?
  - 87% “Definitely Agree”
- Select your ideal advance notification period for a suspected delay of at least 15+ minutes.
  - 83% “2-3 hours”
- Informs the project aim of **detecting 15+ minute delays ~2-3 hours before they happen**

What matters to our passengers?

False Positive Result

False Negative Result

How do we quantify the quality of our model?

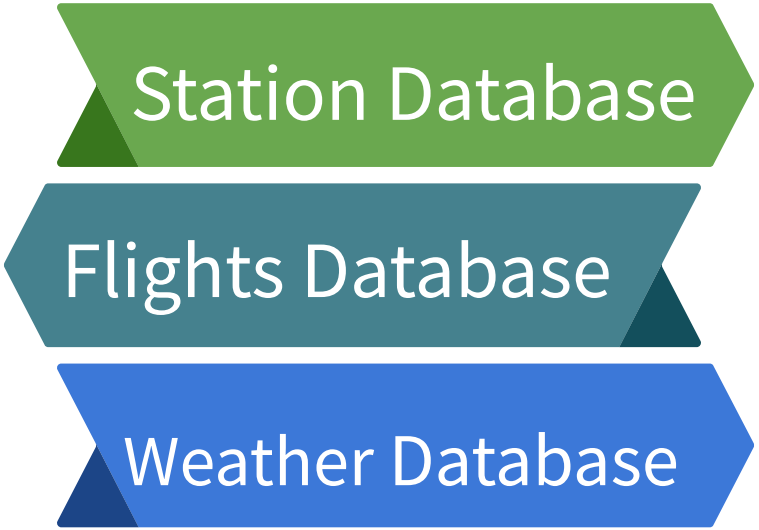
Precision:
Of the flights we say are delayed, what percentage actually are?

Recall:
What percentage of delays do we correctly identify?

F-0.5 Score:
The higher the score, the more likely we can rely on our delay predictions, even at the cost of missing some delays.

Data Sources

Airline performance data from the US Department of Transportation (on time/delayed/cancelled flights, flight IDs, origins/destinations)



Attributes identifying airport locations, other codes (IATA) relevant to identification of an airport from the OurAirports database.

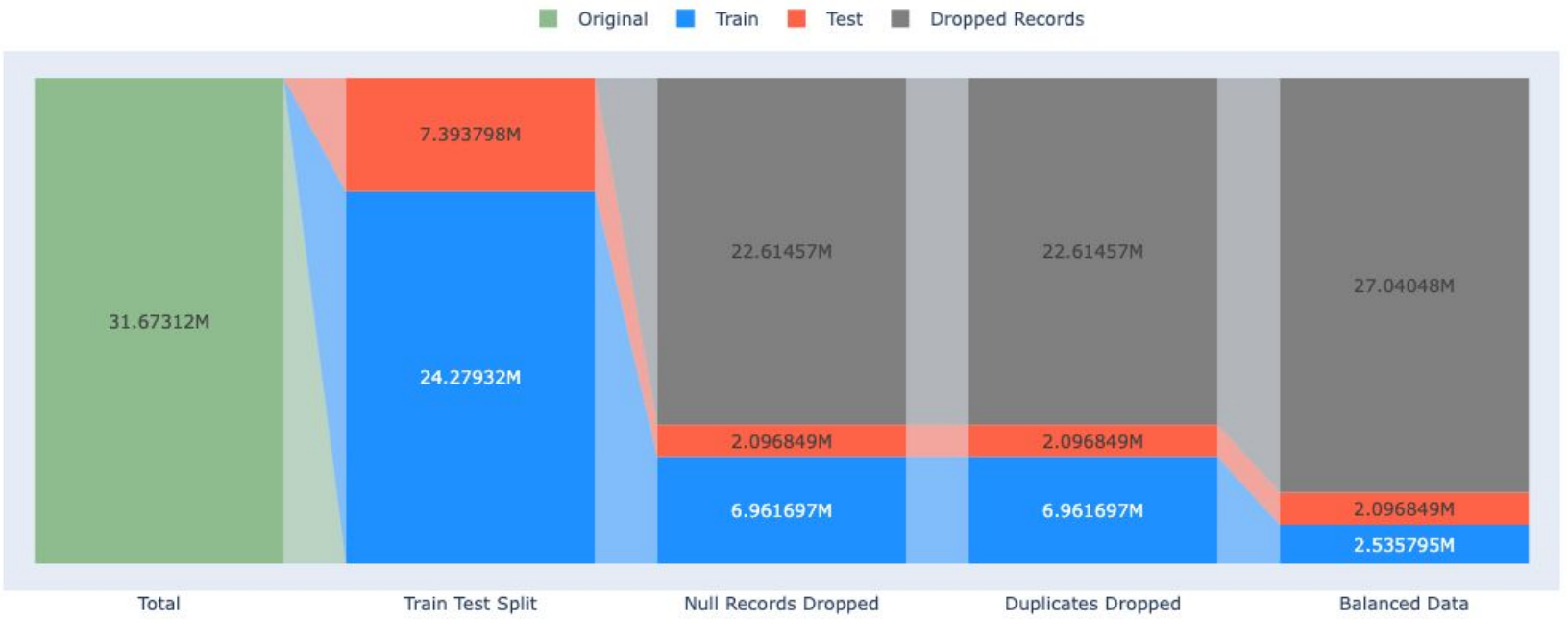
Daily weather data from the US National Oceanic and Atmospheric Administration (temperature, air pressure, precipitation type and level)

2.

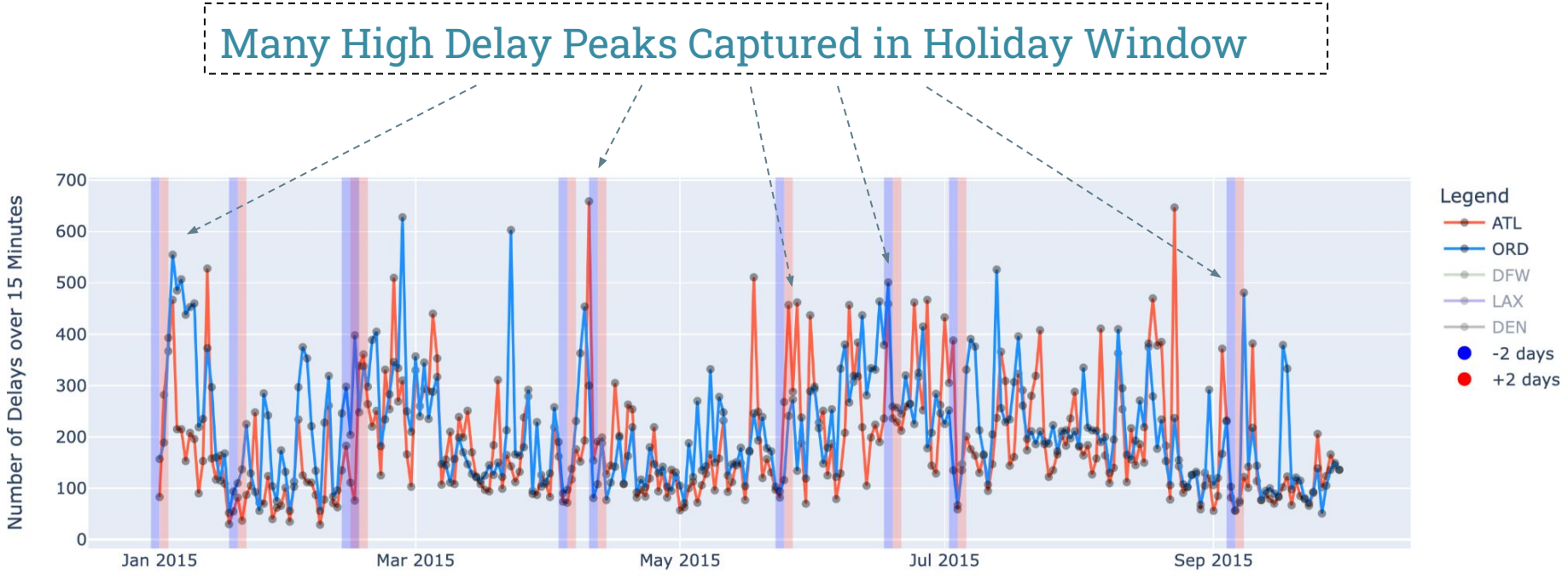
EDA and Feature Selection

What variables do we anticipate most contribute to delays?

Dataset Size



Feature Selection - Federal Holiday Window

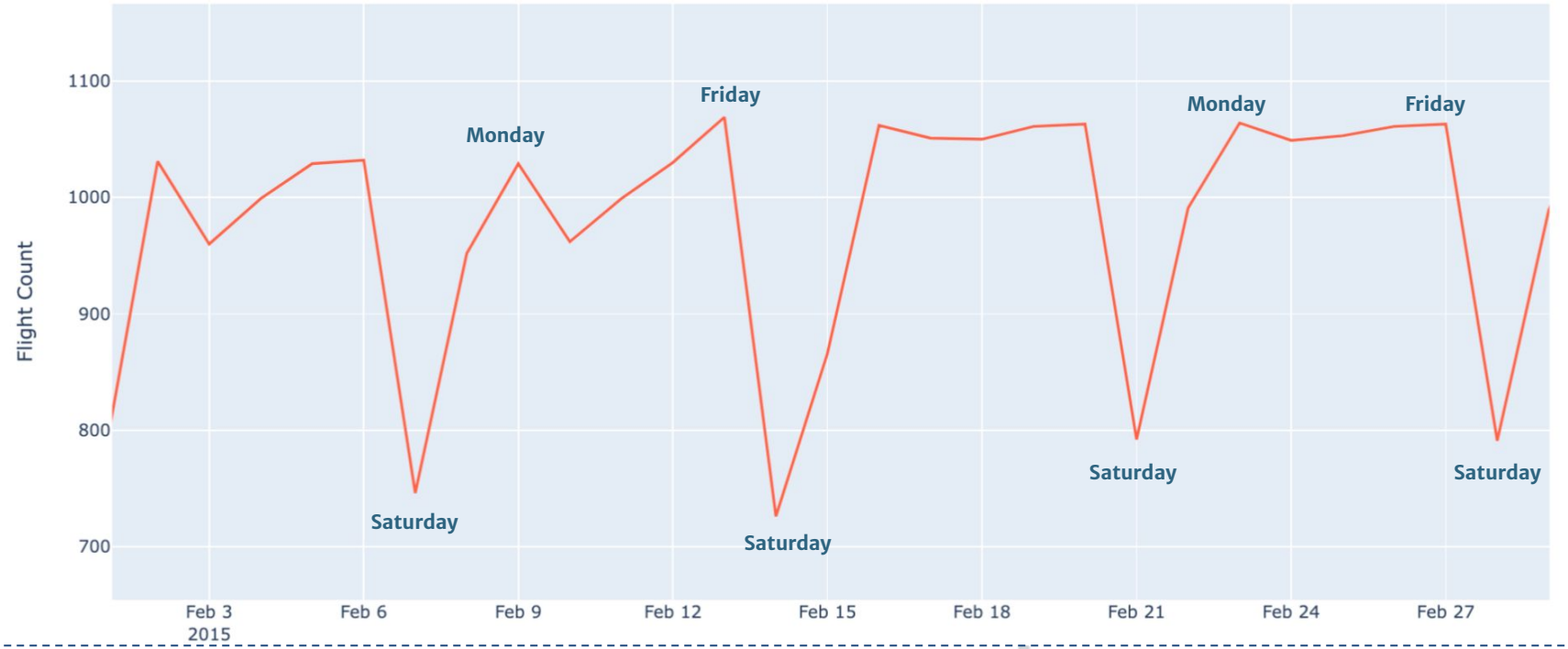


Federal Holiday Window - Example

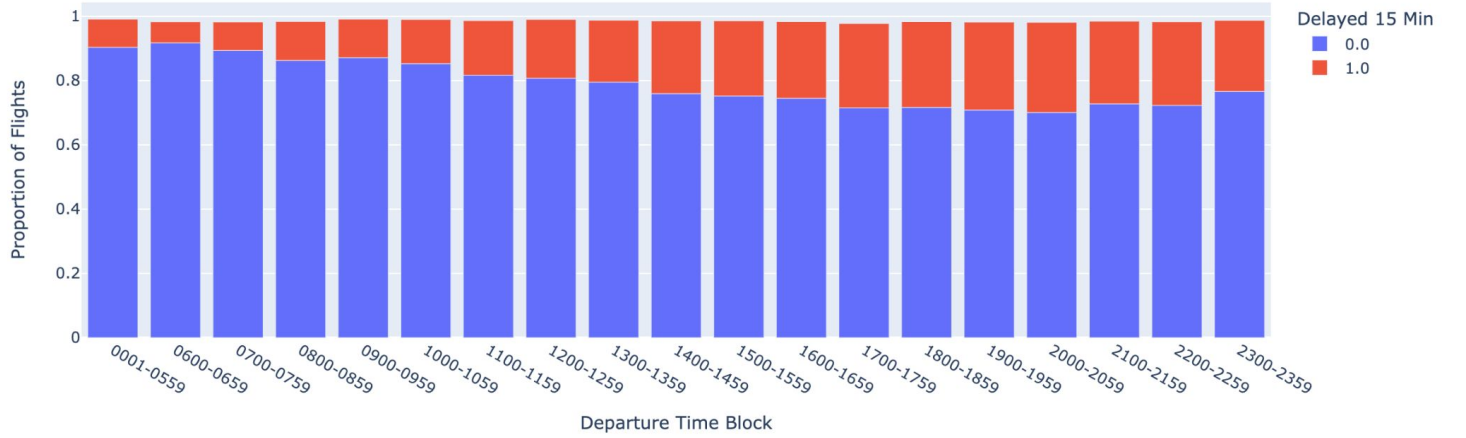
Juneteenth

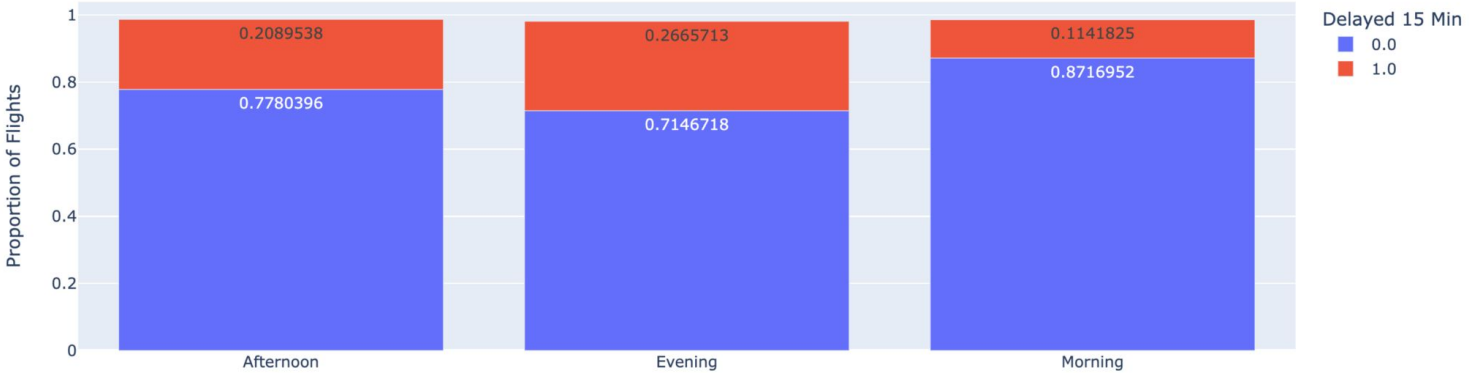
Fourth of July

Feature Selection - Day of the Week



Feature Selection - Time of Day

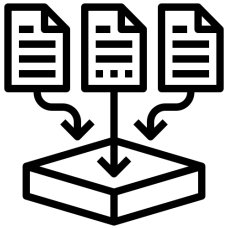




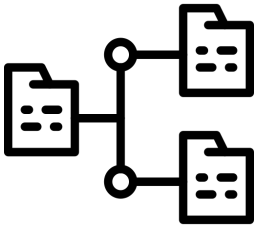
3.
Model Pipeline

How did we go about delivering the best predictions possible?

Model Pipeline



Data Consolidation



Time Series Data Split

EDA & Feature Selection

Model Building & Tuning

Best Model Selection

4.

Model Development

The models we built and reasons we chose them.

Models Overview

- Logistic Regression
  - Simple and efficient model, serving as our baseline model.
- Decision Tree Classification
  - A more complex model capable of extracting nuanced relationships to a tree-like structure and capture non-linear relationships.
- XGBoost
  - A ensemble model of Decision Tree Classifiers, iteratively building better trees to address bias and variance.
- Multilayer Perceptron
  - Most advanced architecture implemented, consisting of neural network capable of capturing the more intricate and non-linear patterns in our data.

5.

Results and Discussion

What models performed well? How do we interpret these results?

5 Year Models Results (macro)

Model	Precision	Recall	F_1	$F_{0.5}$
Logistic Regression (Train)	0.6133	0.6410	0.6268	0.6186
Logistic Regression (Test)	0.2667	0.6350	0.3756	0.3017
Decision Tree (Train)	0.5751	0.7588	0.6543	0.6043
Decision Tree (Test)	0.2388	0.7457	0.3617	0.2763
Multilayer Perceptron 1 Train	**0.6315**	**0.6034**	**0.6171**	**0.6257**
Multilayer Perceptron 1 Test	**0.2818**	**0.5918**	**0.3148**	**0.3148**
Multilayer Perceptron 2 Train	0.6302	0.6193	0.6247	0.6280
Multilayer Perceptron 2 Test	0.2808	0.6033	0.3832	0.3144
3 year XGBoost Train	0.6114	0.6654	0.6372	0.6214
3 year XGBoost Test	0.2587	0.6173	0.3646	0.2927

Results Overview

- Multilayer Perceptron performed the best out of all the models
- Unweighted Results
  - 28.18% overall precision on test set
  - 31.48% F-0.5 score test set
- Confirms our Phase 2 hypothesis that the dataset was nonlinear

6.

Limitations and Future Directions

Where do we go from here?

Limitations / Discussion

- F-0.5 is brutal metric for this form of binary classification with such skew in test labels. Weighted metrics perform much better, but do not pursue the true task of limiting false positives.
  - We feel however, that F(β>1) is doing a disservice to customers if they're interacting with our model.
- Our model seems to overfit, specifically for precision in training.
- Going forward, we will need to implement some form of early stoppage to achieve better generalizability.

Thanks!

Any questions?

You can find us at:

Nathan - https://www.linkedin.com/in/nathanarias/

Thomas - https://www.linkedin.com/in/data-sci-thomas-dolan/

I-Hsiu - https://www.linkedin.com/in/ihsiukao/

Maegan - https://www.linkedin.com/in/maegan-k-8369521a5/