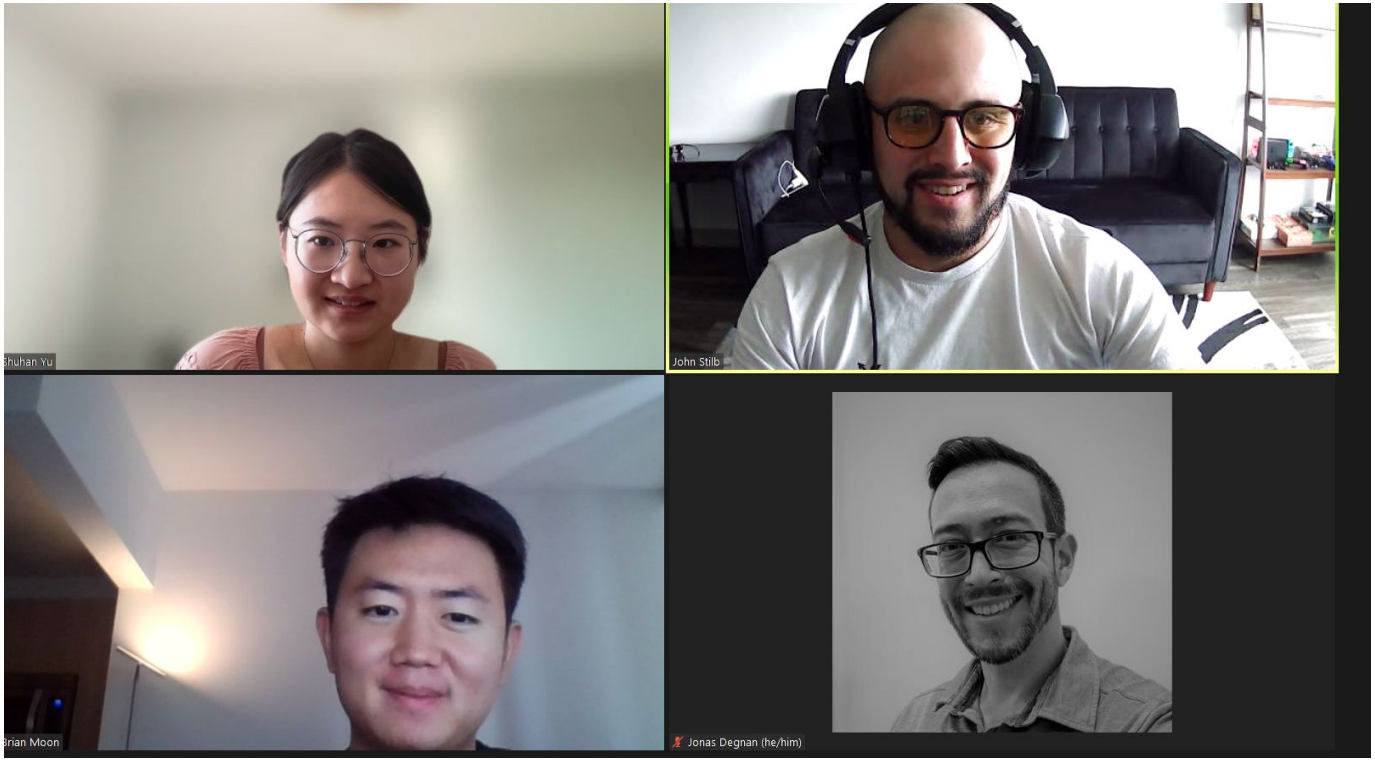
Phase 4 Update: Completed Work, Discussion, & Next Steps

W261 Section 1 Group 2
Brian Moon, John Stilb, Jonas Degnan, and Shuhan Yu

Berkeley
UNIVERSITY OF CALIFORNIA



Berkeley
UNIVERSITY OF CALIFORNIA

Outline

1. Phase Update
2. Feature Engineering
    a. EDA and Feature Selection
3. Modeling Pipelines
4. Results & Discussion
5. Conclusion

Phase Update: Abstract

Our team focused on key shortcomings and discoveries made during the prior phase based on our gap analysis, observations, and instructor feedback.

During this phase our team:

- Improved best model F_1 by **168%**.
- Addressed unbalanced training data.
- Engineered **6** more prognostic features.
- Improved data and modeling pipeline design and workflows.

Phase Update: Description

Objective

- Prediction variable: delayed bivariate (delayed/diverted/cancelled)

Evaluation metrics

- recall, precision, and F_1

Feature engineering

- *Natural disasters/extreme weather*
- *Time-based*
- *Event-based*
- *Graph-based*
- *Derived*

Modeling and Pipeline

- Loss functions: log, hinge, gini impurity
- Regularization: elastic net
- Tuned parameters: regParam, maxDepth, numTrees
- *Blocking cross validation*
- *Ensemble voting model selection*

Feature Engineering

Explored Features:

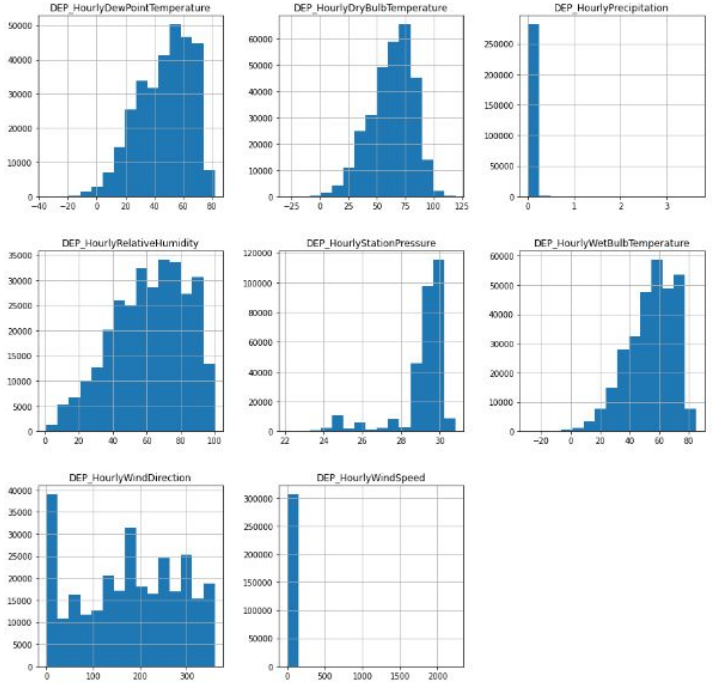
1. Natural disasters/extreme weather
    a. Extreme weather (ice crystals, snow, hail, fog, smoke, storm, or haze)
    b. Icy runway (below freezing point & precipitation >0 in prior six hours)
2. Time-based
    a. Prior extreme weather
    b. Average delay flag over different time windows
3. Event-based
    a. Holiday
    b. Weekends
4. Graph-based
    a. PageRank by airport
5. Size of the airline and airtime

EDA on Explored Features:

A. Correlation to DELAY_FLAG
B. Correlation between explanatory variables
C. Average “delays” per each category, if a categorical variable

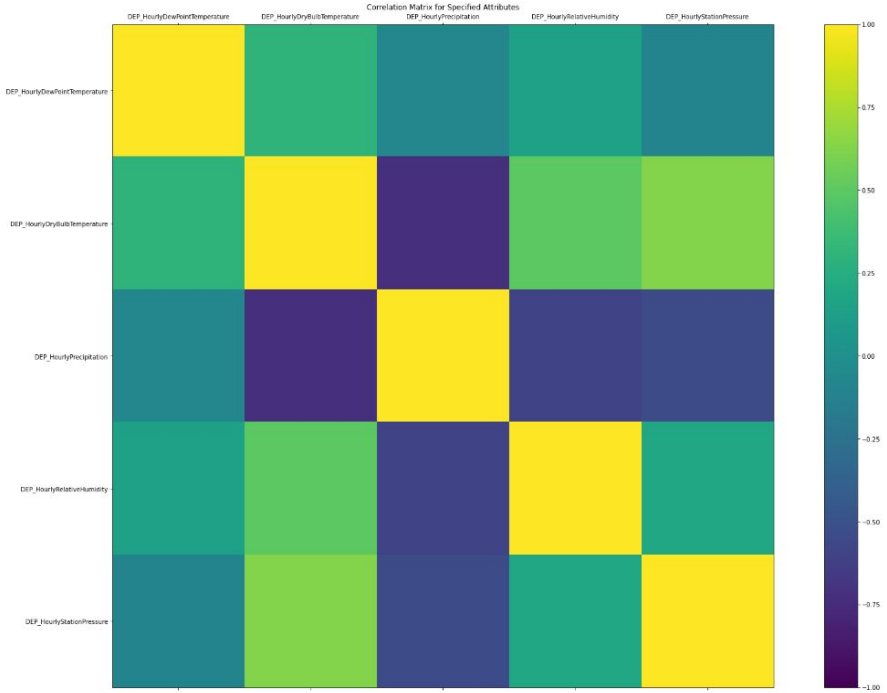
Distribution of Key Explanatory Variables

- Most variables are left-skewed
- Some variables seem to have extreme outliers



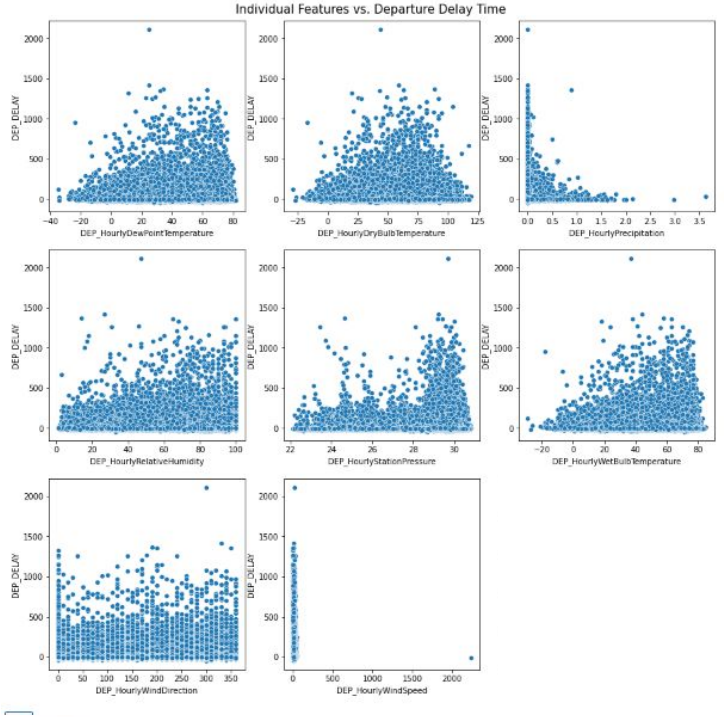
Correlations between Explanatory Variables

- Some explanatory variables have strong negative/positive correlations between each other
- May need to adjust in feature selection/engineering

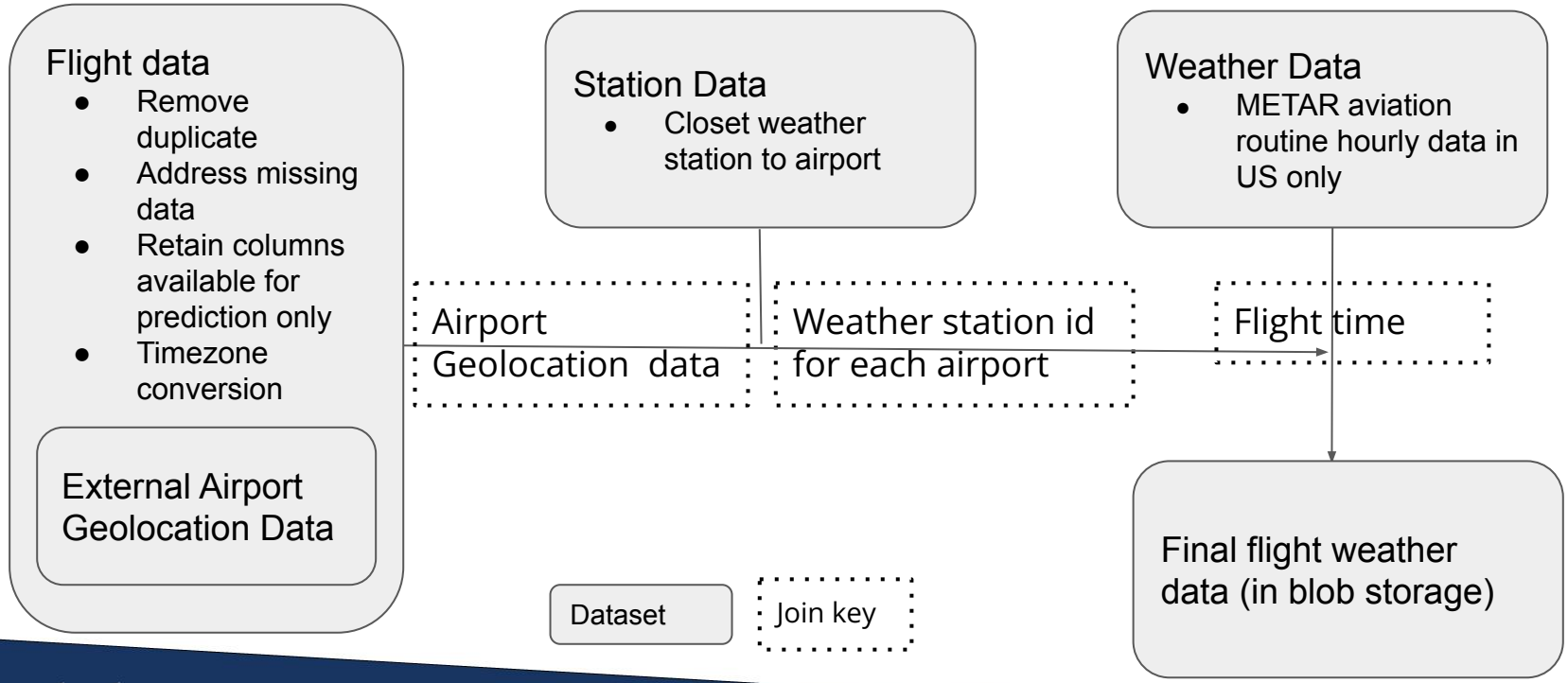


Delay Time vs. Key Explanatory Variables

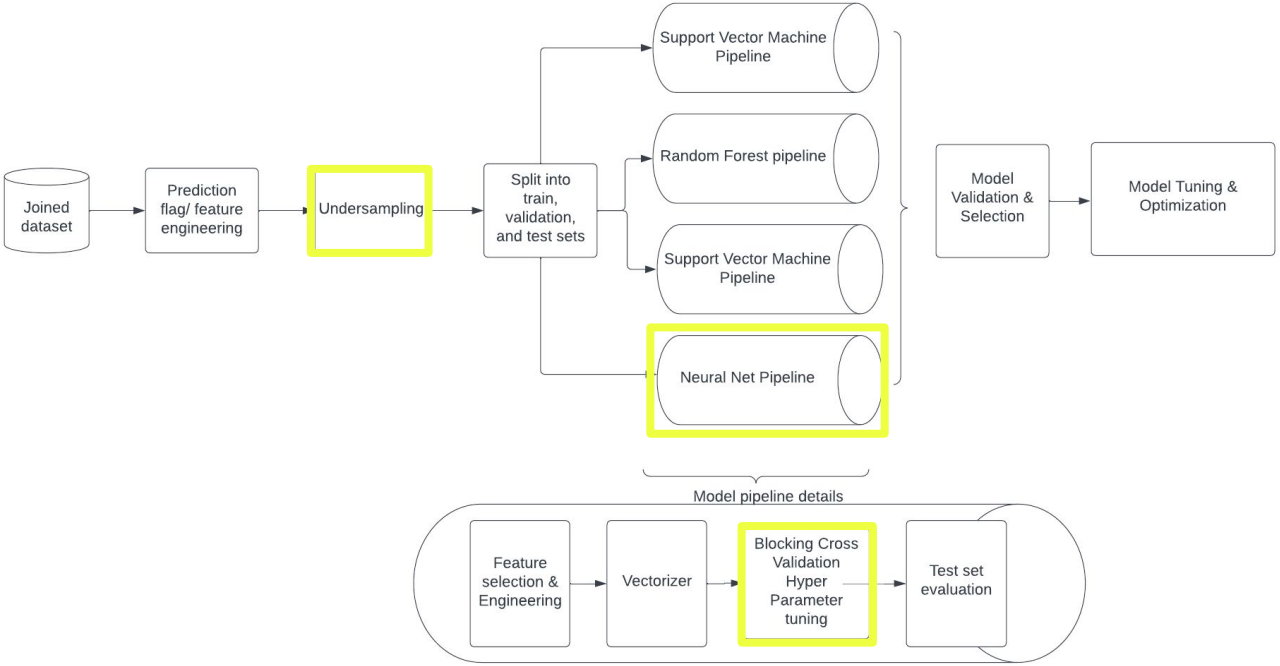
- Overall, more delays in more extreme values, yet not that informative as it does not account for frequency



Joining the data



Modeling Pipelines



Results and Discussion

	#1 Model: Random Forest	#2 Model: LSVC	#3 Model: Log Reg	Prev Best Model: Random Forest
Features	Key Engineered Features	Key Engineered Features	Key Engineered Features	Key Features + Rolling Avg Delays
Metrics	F1, Precision, Recall	F1, Precision, Recall	F1, Precision, Recall	F1, Precision, Recall
Performance	Precision = 0.339 Recall = 0.586 F1 = 0.429	Precision = 0.322 Recall = 0.537 F1 = 0.402	Precision = 0.314 Recall = 0.558 F1 = 0.402	Precision = 0.927 Recall = 0.087 F1 = 0.160

Conclusion and Next Steps

Open Issues & Problems

Backup Slides