Predicting Flight Delays through Machine Learning Classifiers at Scale Phase IV Update

W261 Fall 2022 Section 5 Group 4:
Nathan Chiu, Dominic Lim, Raul Merino, Javier Rondon

Airlines should implement machine learning at scale to better predict flight delays for resource allocation / customer service purposes.

Problem

20% of flights are delayed by more than 15 minutes, costing tens of billions of dollars

Data

Flights, weather, weather station, and airport codes from government agencies

Strategy

Optimize model selection for F2 performance and minimal run time

Models

BASELINE	Logistic Regression for Simplicity
DECISION TREES	Efficiency and Performance at Scale
RANDOM FORESTS	Collected Decision Trees
GRADIENT BOOSTED TREES	Accelerating Learning Rate
ENSEMBLE	Combining Model Predictions

Modeling Steps

- Feature Engineering
- Model Pipelines
- Hyperparameter Tuning
- Run Experiments & Gap Analysis

Project Outcomes

- **Lift in F2 Score** driving more confidence in airlines' business decisions
- **Faster runtimes** enabling timelier insights on flight delays

Feature Engineering

We created new features to boost the predictive power of our models

Feature Name	Description
Previous Flight Delay	Airlines have a finite number of aircrafts, so each aircraft has a route that it follows every day, going from airport to airport often involving back to back scheduled flights. An earlier delay may affect subsequent flights for the same aircraft
Pagerank Features	PageRank describes an airport's importance and influence, which can describe how delays are spread throughout a network of airports.
Delay States	The delay state represents the network's delay patterns at a point in time
Weather Features	The categorical features indicate the presence of weather related to flight delays such as thunderstorms, snow, fog and ice
Average Airport Delay	We created a feature for the percentage of flights that are delayed in a given time window
Airport Capacity	The ratio of actual flights that depart over scheduled flights out of an airport

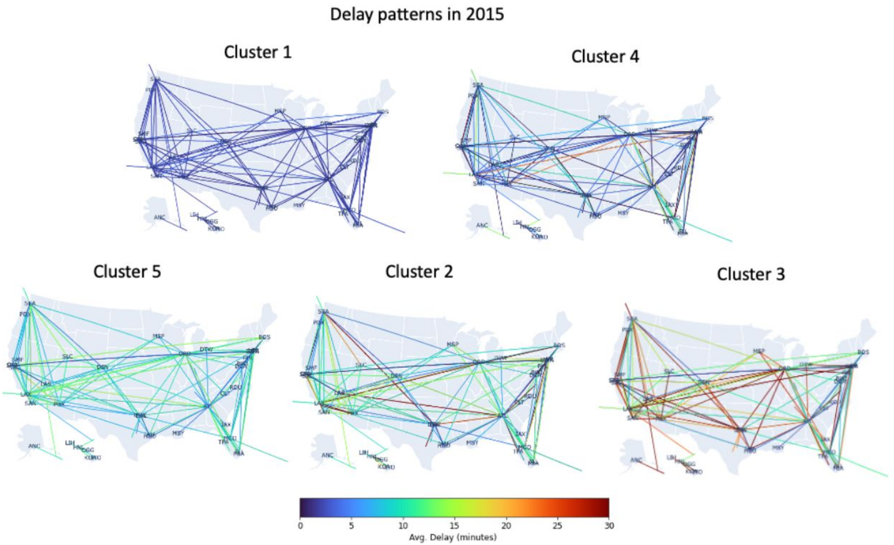
Feature Engineering EDA

We conducted an exploratory data analysis of the newly engineered features, focusing on understanding the features' distribution, scale, and range of values

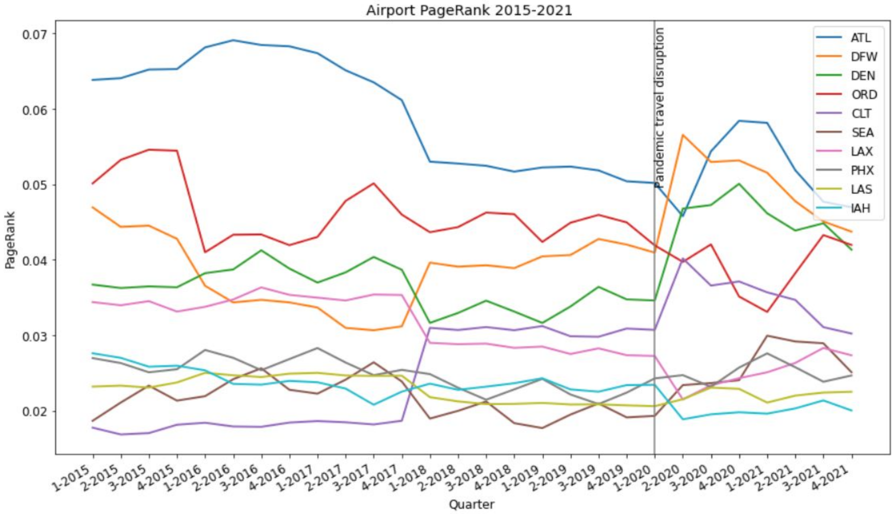
Previous Delay

- After an initial analysis, we saw that this had a relatively **high correlation (> 0.3)** with the current's flight delay

Delay State



Airport PageRank



- PageRank shows that the most important airports are **ATL and DFW** and changes in rank across time
- In the delay state cluster with the most delays stem from flights that involve **DFW, ORD and LAX**

Models Pipeline

We focused on generalizing functions involved in the pipeline to make it easier to adjust parameters and run a multitude of experiments

Feature Selection

We began by running decision tree models with different categories of features:

- Weather Features
- Airport Capacity (QRN)
- Airport PageRank
- Clustered Delay States
- Previous Flight Feature (based on Tail Number)
- Other Flight Features (Airline Carrier, Seasonality)

Hyperparameter Tuning

Once features were selected, we experimented with combinations of parameters against cross validation data

- Decision Trees / MLP: VectorAssembler, MinMaxScaler
- Decision Tree Loss Function: Gini Impurity

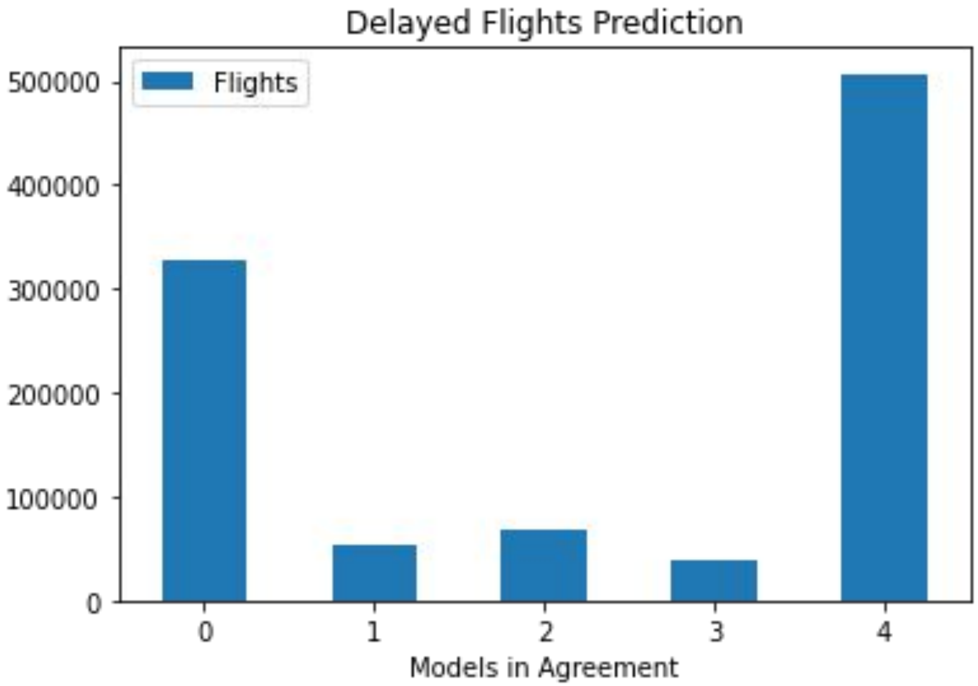
Model Selection

Once we selected the best hyperparameters, we compared the primary metrics like F2 score, precision, and recall across all models:

- Used average F2 score to fit the full train dataset and evaluate the full test dataset

New Model

We also wanted to implement novel approaches including the use of ensemble methods whereby all four models (hyper-parameterized Decision Tree, Random Forest, Gradient Boosted Tree, and Multilayer Perceptron) "vote" on the final prediction



Voting Mechanism

Vote by Majority: The majority prediction of DELAY or NO DELAY

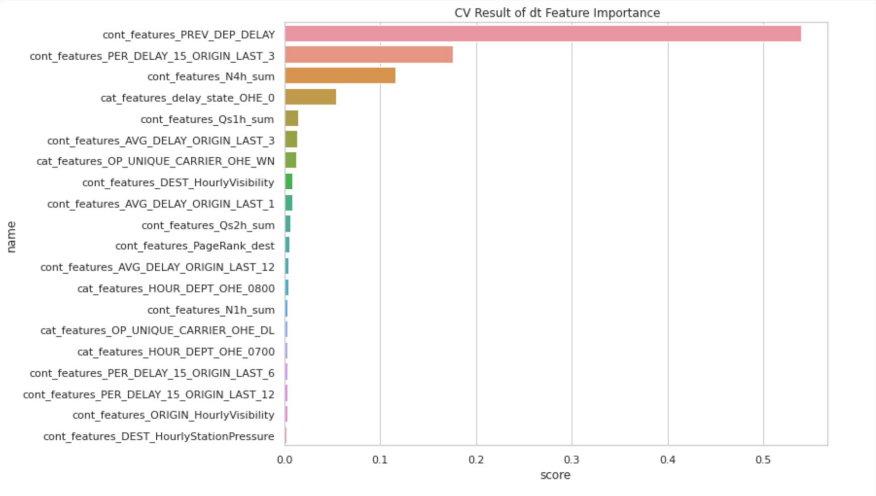
One Positive Voting: If one model suggests delay, predict DELAY

One Negative Voting: If one model suggests no-delay, predict NO DELAY

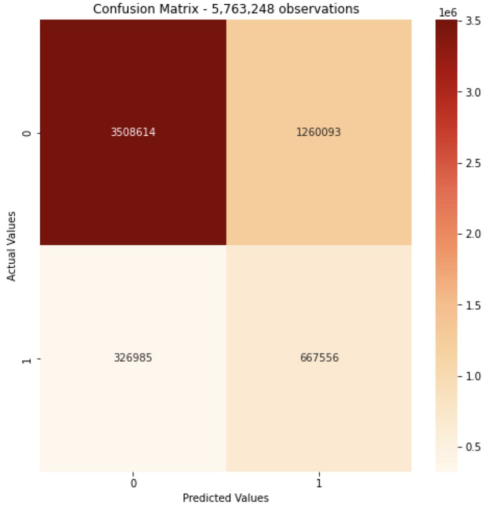
Experimental Results

We compared primary metrics of success like F2 across hyper-parameterized models and the Ensemble models performed the best

Feature Importance



Best Model: Ensemble Confusion Matrix



- The F2, Precision and Recall score range from 54.7%, 41.1%, 61.2% to 55.8%, 36.6% and 64.3%
- **67%** of the delayed flights are correctly classified

Model	Layers	Max Bins	Max Depth	Max Iterations	Number of Trees	Train F2	Train ROC AUC	Train Precision	Train Recall	Test F2	Test ROC AUC	Test Precision	Test Recall
MLP	[44, 44, 2]	-	-	100	-	0.641	0.748	0.716	0.619	0.519	0.755	0.388	0.589
Decision Tree	-	350	10	-	-	0.617	0.765	0.760	0.589	0.540	0.764	0.411	0.586
Gradient Boosted Tree	-	100	10	6	-	0.630	0.772	0.756	0.605	0.546	0.771	0.405	0.599
Random Forest	-	50	10	-	100	0.642	0.765	0.737	0.622	0.547	0.765	0.384	0.612
Ensemble	-	-	-	-	-	-	-	-	-	0.558	-	0.366	0.643

Conclusion

Airlines should implement machine learning at scale to better predict flight delays for resource allocation / customer service purposes.

Problem	20% of flights are delayed by more than 15 minutes, costing tens of billions of dollars
Data	Flights, weather, weather station, and airport codes from government agencies
Strategy	**Select Ensemble Model for F2 performance and minimal run time**

Models

BASELINE	Logistic Regression for Simplicity
DECISION TREES	Efficiency and Performance at Scale
RANDOM FORESTS	Collected Decision Trees
GRADIENT BOOSTED TREES	Accelerating Learning Rate
ENSEMBLE	Combining Model Predictions

Modeling Steps

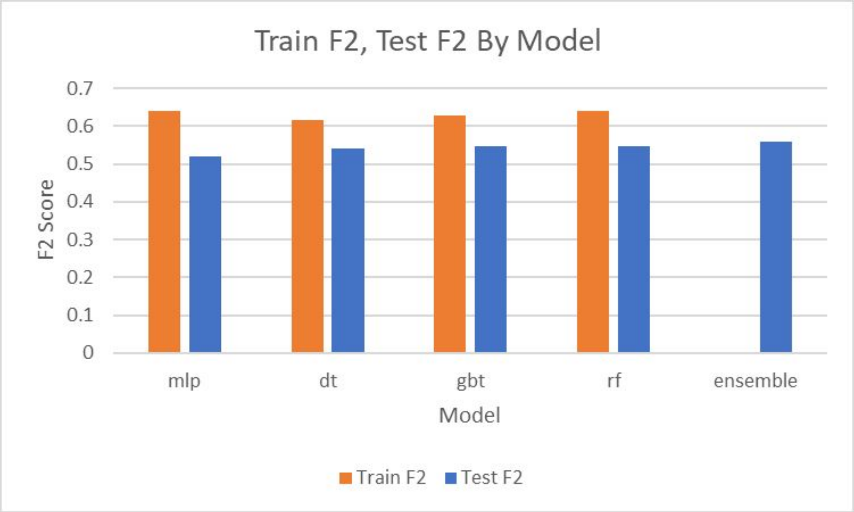
- Feature Engineering
- Model Pipelines
- Hyperparameter Tuning
- Run Experiments & Gap Analysis

Project Outcomes

- **F2 Score of** .558, nearly 5x the baseline
- **Fast runtime** of two minutes

Phase IV: Appendix

Appendix



Results & Discussion

- The ensemble model followed by the random forest performed the best on the test dataset with F2 scores of .558 and .547 respectively