Domestic Flight Delay Predictions

W261 Final Project:

Beijing Wu - Grace Lee - Shivangi Pandey - Sybil Santos-Burgan

BOARDING PASS

- **FLIGHT**
  W261
- **GATE**
  Summer 2022
- **SEAT**
  Section 4

FP_Section4_Group1

Agenda

- BUSINESS CASE
- DATASETS
- EDA
- FEATURE ENGINEERING
- MODELING
- RESULTS
- CHALLENGES and FUTURE OPPORTUNITIES

Business Case

Project Objective:

Can we predict departure delays 2 hours before scheduled CRS time?

Why do we care?

Delays are rising as the flight industry is struggling to keep up to higher travel demand post-COVID

Main Client:

Assist **Airports** in predicting delays to help them plan operations

Data

Flight

Domestic flight and airport data between 2015 - 2021

From the US Department of Transportation

Weather

Hourly weather data between 2015 - 2021

From National Oceanic and Atmospheric Administration

Station

Weather station and neighboring station data along with distance between station information.

From the US Department of Transportation

Supplemental Data

Airline Names

Airline name description by IATA code

From the Bureau of Transportation Statistics

Holiday

Holiday travel seasons as defined by the air travel industry

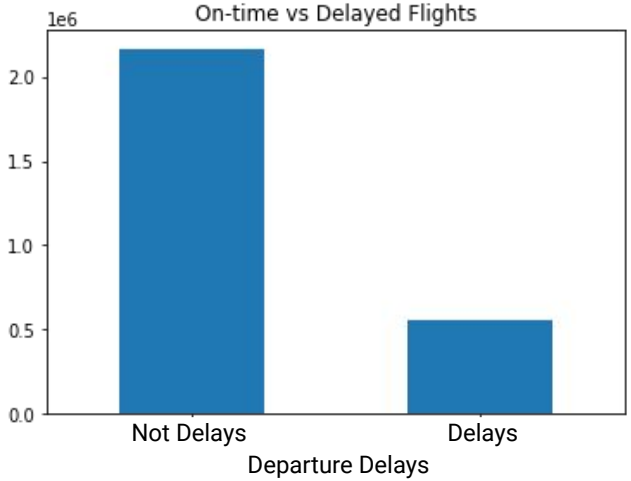
From the Bureau of Transportation Statistics

Carrier Category

Carriers by their operation cost structures - low-cost carrier (LCC), regional airlines, and legacy (major airlines)

From ICAO and Airline Pilot Central

EDA



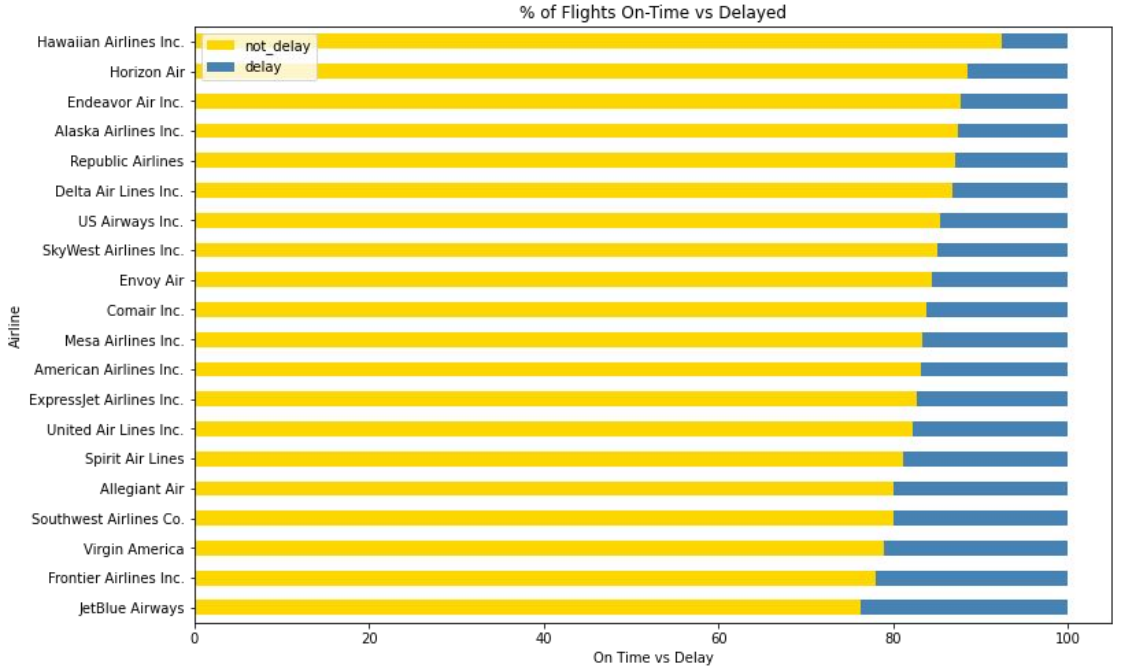
“Missing” delays indicate cancellations
⟶ We consider these indefinite delays

~30%
of flights depart delayed

EDA

Low Cost Carriers tend to have higher % of flights delayed

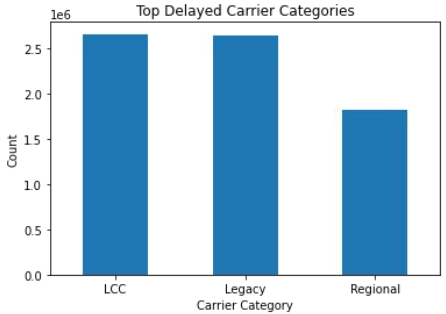
Low Cost Carriers



Feature Engineering

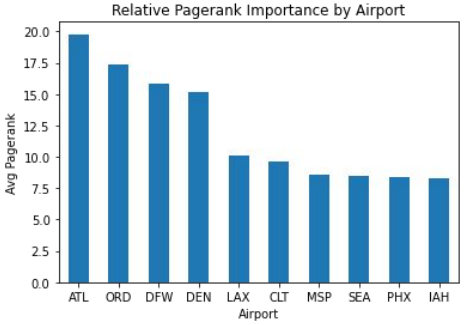
Carrier Category

Low cost carriers have higher % delayed → created a carrier category variable to account for trends by airline type



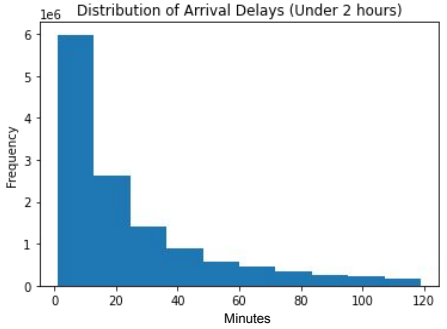
Pagerank of Origin Airport

Airports can have a range of activity → quantify activity and importance of departing airports via pagerank



Prior Flight's Late Arrival

"Chain" effect of late arrivals impacting future departures → create indicator if the plane previously arrived 30+ min late



Model Evaluation

Metrics for Success

We assume airports want to predict more delays, even if predictions are false, to better plan operations

01 

F2

A weighted mean of precision and recall, emphasizing minimization of false negatives

02 

MATTHEWS CORRELATION COEFFICIENT (MCC)

Representation of levels of true and false predictions, weighting the proportion of delayed flights

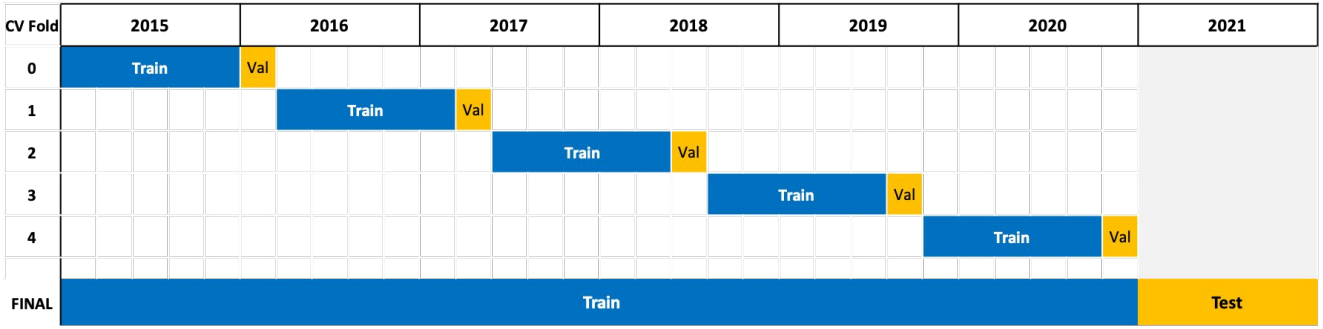
03 

BALANCED ACCURACY

Accuracy, taking into account imbalanced delay data

Model Pipeline

5-Fold Rolling Window CV



Class Imbalance

Downsample majority (not delayed) class

Add Features

Tune Hyperparameters

Grid Search
Random Search

Modeling

Algorithms

Logistic Regression

Estimates probability of event occuring

Decision Tree

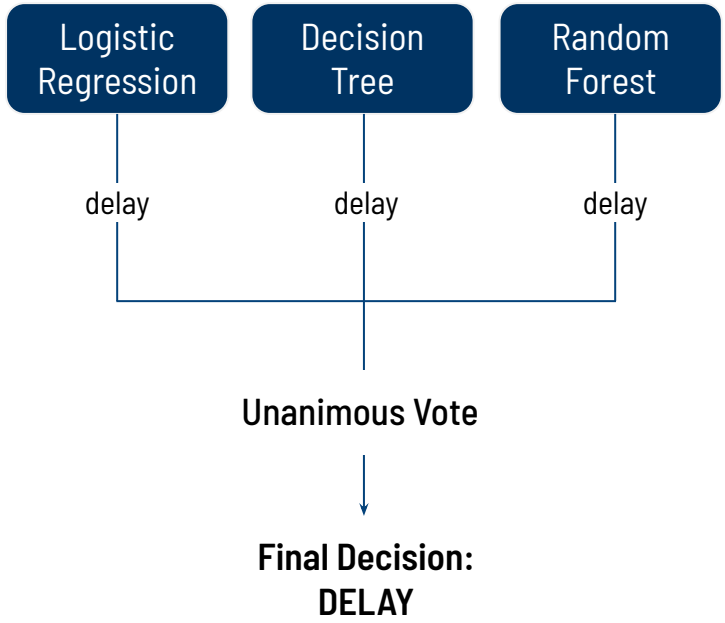
Model of potential outcomes split based on different attributes

Random Forest

Collection of decision trees

Ensemble

Combines predictions from multiple other models



Model Performance

Logistic Regression

Actual Value \ Predicted Value	0	1
0	4,914,771 (0.99)	55,981 (0.01)
1	470,076 (0.46)	554,569 (0.54)

F2: 0.5888

MCC: 0.6596

BA: 0.7650

Decision Tree

Actual Value \ Predicted Value	0	1
0	4,914,777 (0.99)	55,975 (0.01)
1	470,282 (0.46)	554,363 (0.54)

F2: 0.5886

MCC: 0.6595

BA: 0.7649

Random Forest

Actual Value \ Predicted Value	0	1
0	4,914,511 (0.99)	56,241 (0.01)
1	470,020 (0.46)	554,625 (0.54)

F2: 0.5888

MCC: 0.6595

BA: 0.7650

Ensemble

Actual Value \ Predicted Value	0	1
0	3,966,496 (0.80)	1,004,256 (0.20)
1	378,340 (0.37)	646,305 (0.63)

F2: 0.5607

MCC: 0.3596

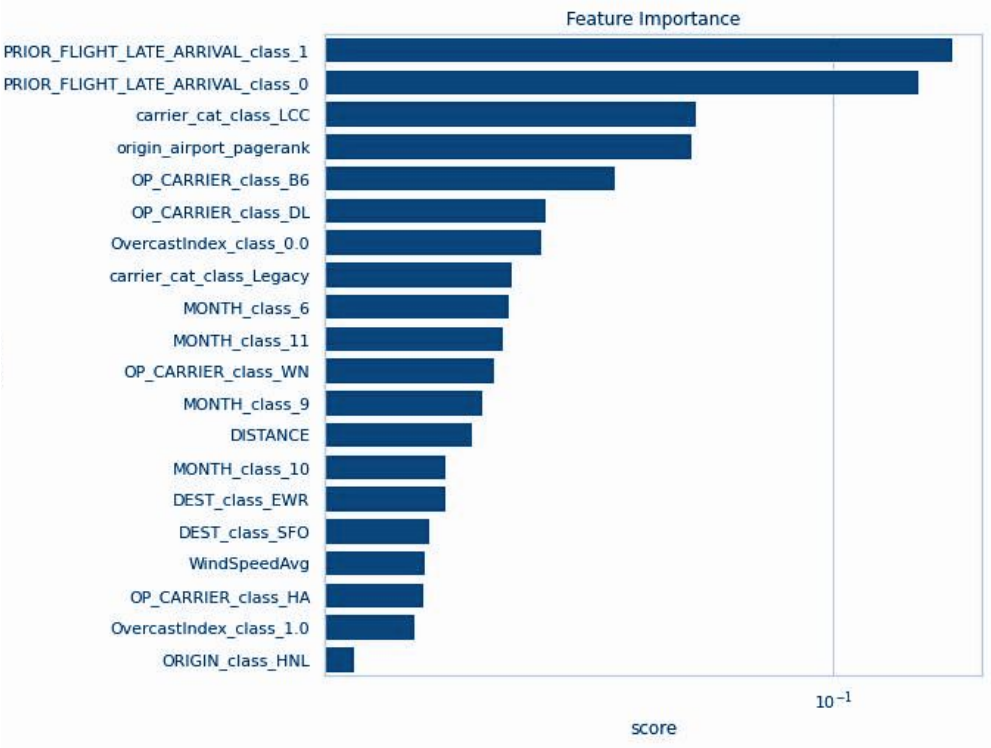
BA: 0.7134

Baseline (Logistic Regression)
F2: 0.227

BA = Balanced Accuracy

Feature Importance

- Prior flight late arrival
- Low Cost Carrier (LCC):
  JetBlue, Southwest
- Airport Importance
  Atlanta, Chicago O'Hare
- Overcast Weather



Challenges

- **Time: Joins and Cross Validation**
- **Imbalanced Data:** limitations on seasonal and event-based information (such as holiday and natural disasters)
- **High Dimensionality: Categorical Data**

Future Opportunities

Additional Features

- Additional data sources
- Feature for propagating delay impacts
- International flight information
- Pilot hours per day

Other Techniques

- Weighted Window CV
- Other techniques to combat class imbalance (oversampling, SMOTE)
- Threshold tuning

Additional Models

- XGBoost
- Other ensembling methods using bagging or stacking

Computation

- Leverage Databricks Delta Lake

THANKS!

BOARDING PASS

- **FLIGHT**
  W261
- **GATE**
  Summer 2022
- **SEAT**
  Section 4

FP_Section4_Group1