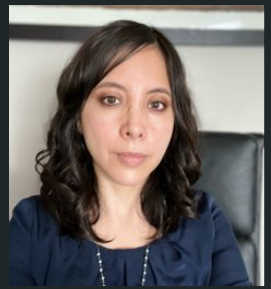
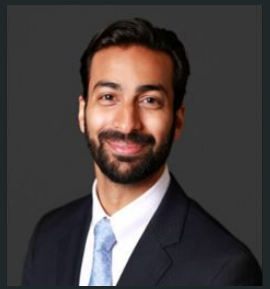
Predicting Flight Delays with Machine Learning Techniques

Team 21
Members: Carla Cortez, Redwan Hussain, Anqi Liu, Murray Stokely

Project Background

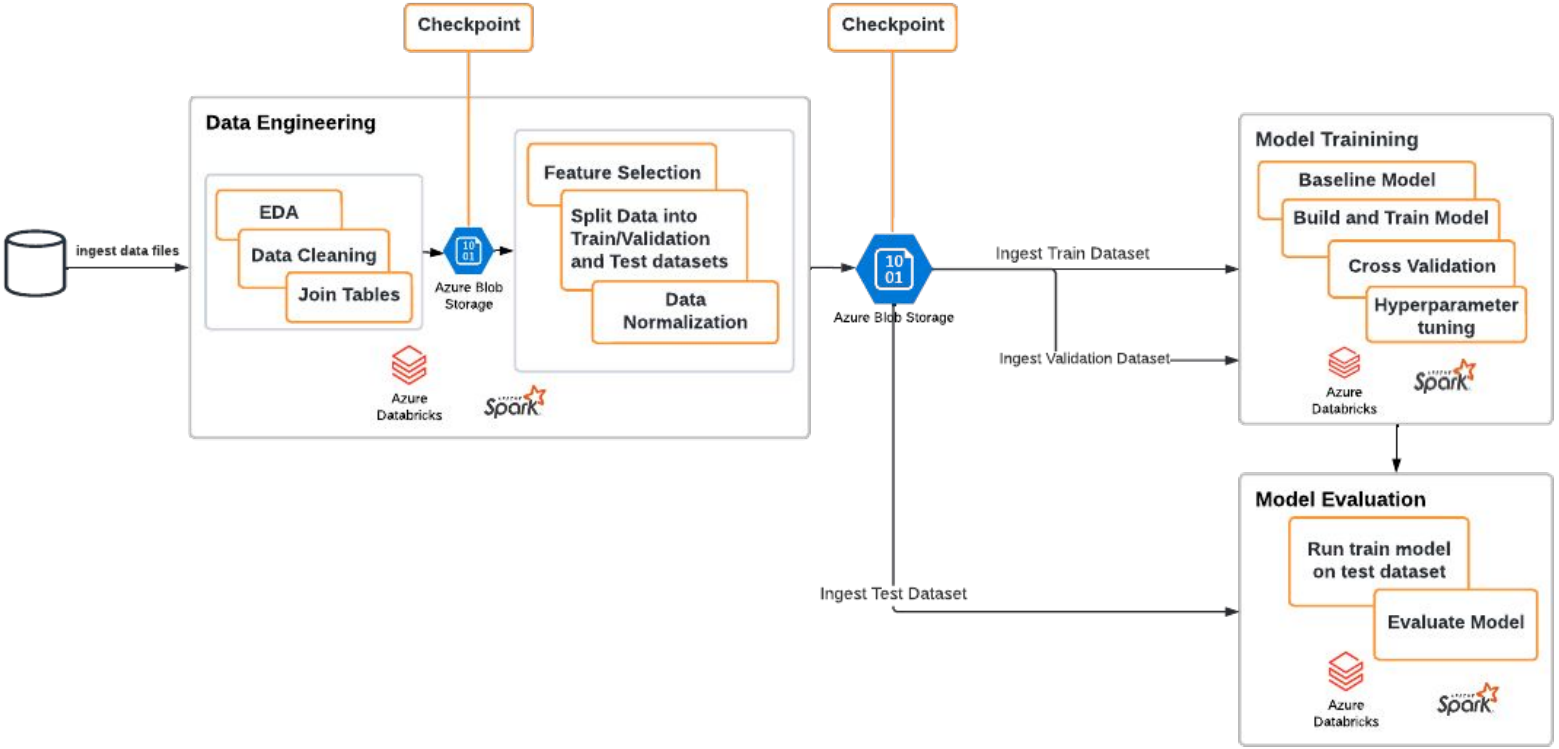
Business Problem

- Airline companies are greatly interested in predicting flight delays to manage financial impact and retain customers
- The ability to predict a delay will help mitigate costs associated with rerouting flights and differentiate against competitors

Proposed Solution: develop classification model to predict delays within 2-hour window using flight and weather data from 2015-2021

- Measure success by ensuring as many customers make their flight
- False Negative: a delay occurs but is undetected and customers miss their flight
- False Positive: a delay does not occur and flights are rerouted, but customers make their flight
- Use precision and recall as evaluation parameters
- F2 score to have more emphasis on recall

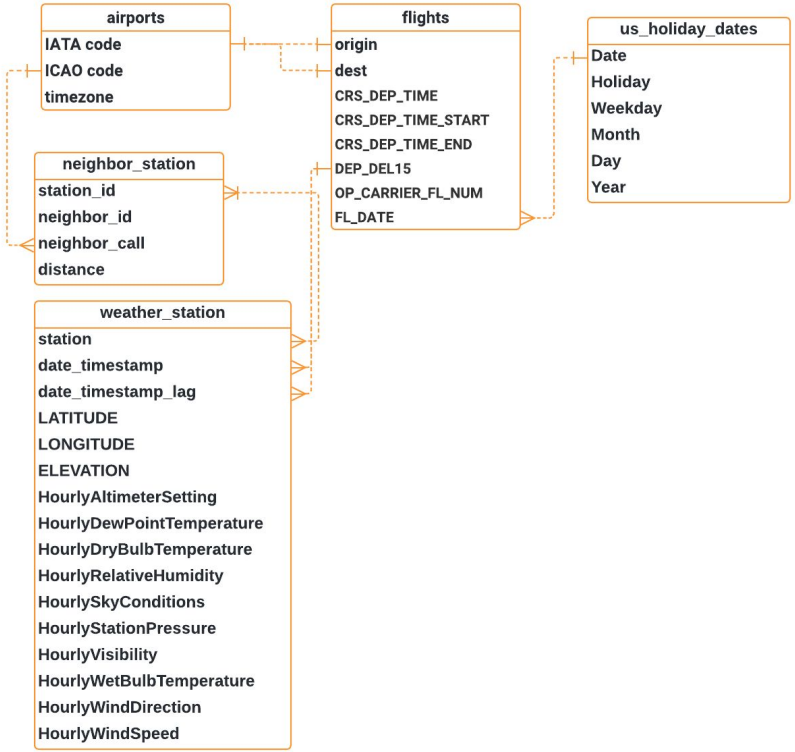
Modeling Pipeline



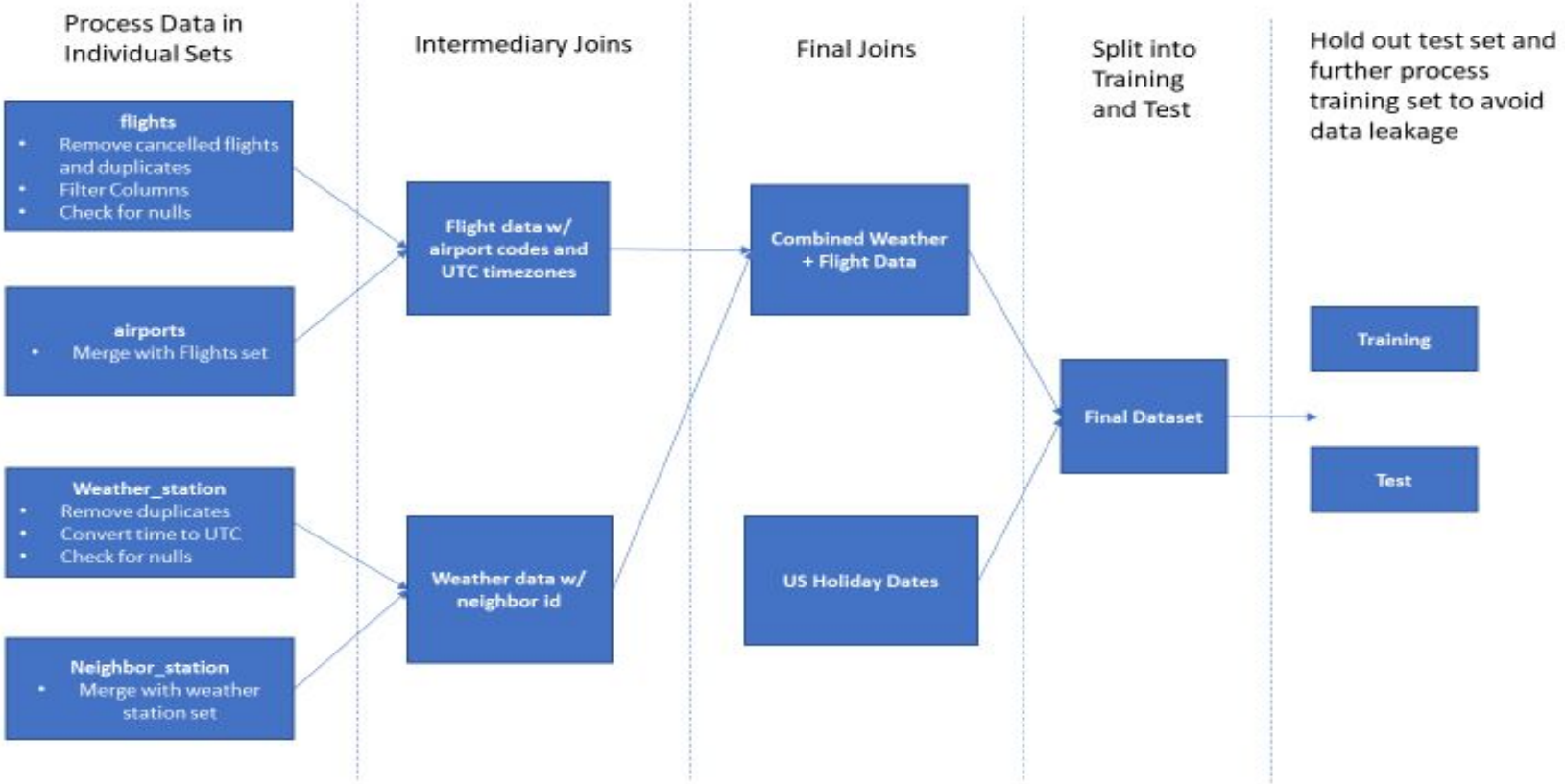
About the Data

Raw Input Data:

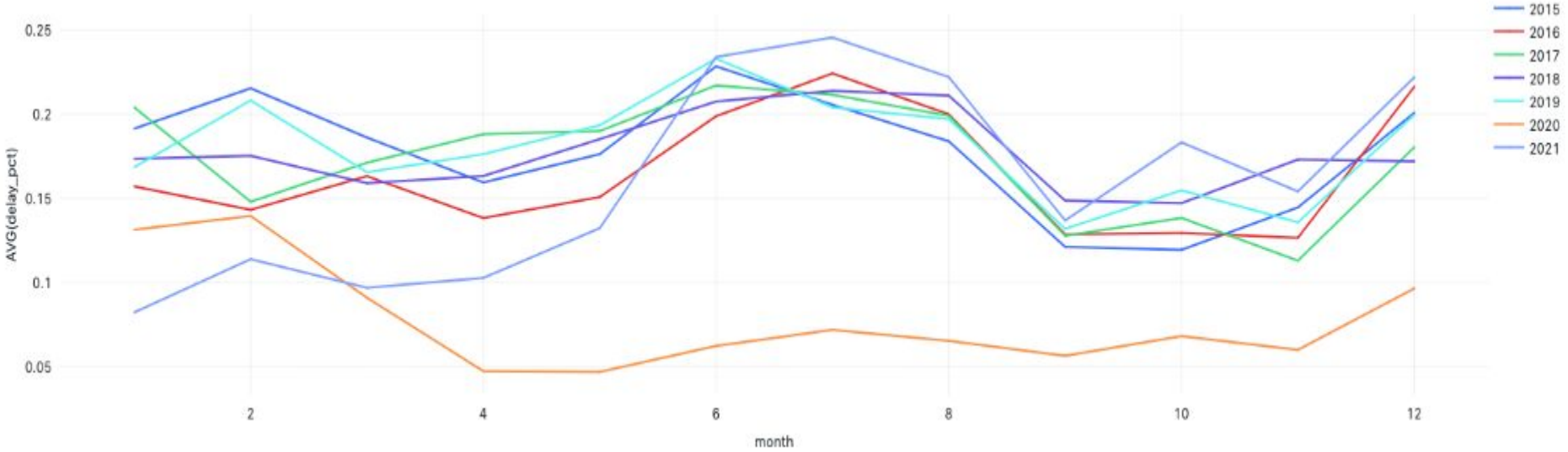
1. US Department of Transportation: Passenger flight on-time performance data
2. National Oceanic Atmospheric Administration: weather dataset
3. IATA ICAO metadata for airport codes and time zones



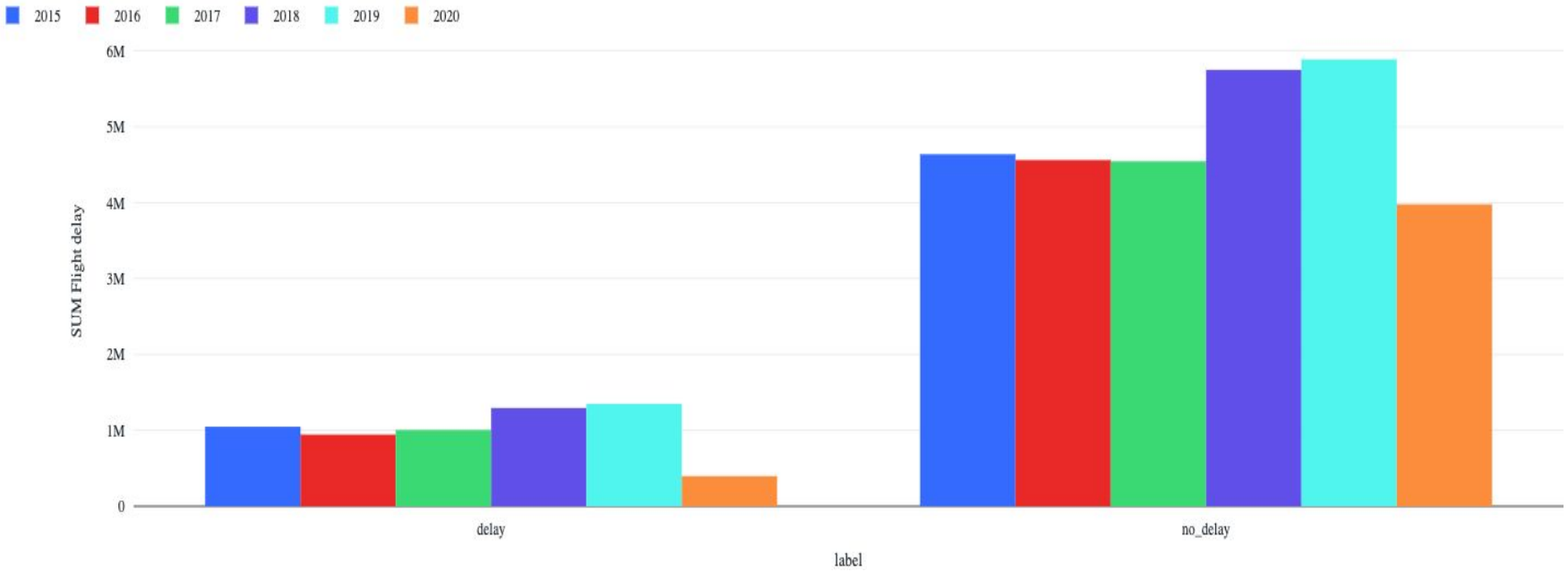
Data Lineage



Data Preprocessing - Seasonality and Outliers

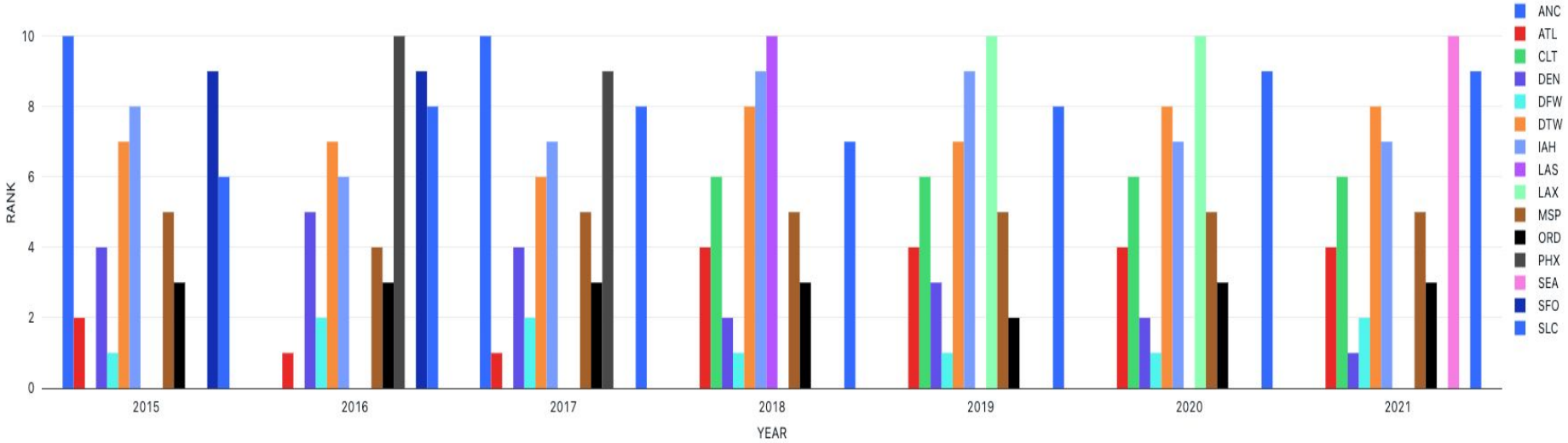


Data Preprocessing - Downsampling



Feature Engineering

- Created holidays feature (event-based) - it shows seasonality in flight volumes
- Created 20 Lag features corresponding to the hourly weather telemetry measured one timestamp before the 2 hours timestamp prior to departure (time-based)
- Used pagerank to create a feature that identifies the traffic per origin airports per year (graph-based)



Data Leakage

- Normalization - Using Mean Only From the training set
- Page Rank Per Year
- Events Unknown Two Hours Before Flight Departure

Feature Selection

- Only well populated variables are kept for further analysis
- Analyzed Pearson correlation between each feature and response variable (DEP_DEL15)
- Hourly wind speed (ws_origin_HourlyWindSpeed) identified as most important feature in terms of correlation

Model Selection

- Baseline: Always Predict Delay
- Logistic Regression
- Random Forest
- Gradient-boosted Tree
- Multilayer Perceptron Classifier - NN

Modeling Experiments

- All experiments were performed using 58 downsampled, encoded, vectorized, min max scaled selected features
- Prepared with one-hot encoding for boolean and label encoding for string variables and VectorAssembler to vectorize all features to prepare them to be used it with all our models.
- Grid-search through the provided Custom Cross Validator for hyperparameter selection
- Evaluate model on the Cross Validation dataset built from the train dataset with pyspark.ml.evaluation.MulticlassClassificationEvaluator
- Experiment with different subsets of 2020 data (include entire year, use Jan-Feb only, exclude year)

Results: Feature Importance from Random Forest

Feature Name	Importance
ws_origin_HourlyWindSpeed*	0.0934
ws_dest_HourlyWindSpeed_lag	0.0911
ws_origin_HourlyWindSpeed_lag	0.0668
MONTH	0.0523
ws_origin_HourlyVisibility	0.0484

*Same result as Pearson correlation between each feature and the response variable

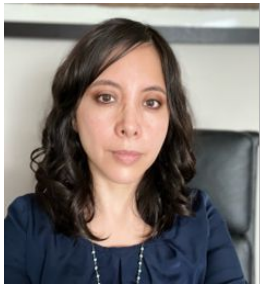
Results: Evaluation Parameters

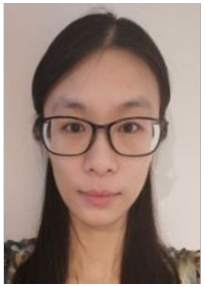
Model	Precision	Recall	F1	F2	Execution Time
Baseline - Always Predict	0.174	1.0	0.296	0.513	N/A
Logistic Regression	0.563	0.594	0.579	0.588	6 mins
Random Forest	0.593	0.560	0.576	0.566	1.44 hours
Gradient-boosted Tree	0.552	0.489	0.518	0.500	20 mins
Multilayer Perceptron	0.563	0.537	0.550	0.542	55 mins

Conclusions and Final Thoughts

- Overall, project was successful - able to predict delays more than 50% of the time
- Gap analysis: many teams that were using precision-recall-F2 scored between 0.5 and 0.6 with a similar number of records, so our results were consistent with others
- Significant improvement once we downsampled the data per year.
- Challenges: limitations due to cluster size
- Recommendation: continue work in a future project with 2022 training data, use ensemble technique

Thank you!






Link to notebook:
https://adb-731998097721284.4.azuredatabricks.net/?o=731998097721284#notebook/1215577238237907/command/1215577238237908