Final Presentation
Team 13

Erin Werner, Ian Dela Cruz, Jae Park, Song Park

The Case for Flight Prediction & Data Sets

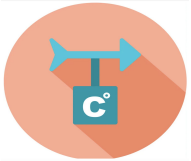
Impact: Increased operational efficiency, customer experience, etc.

Goal: Create a model to predict whether or not a departure delay > 15m will occur for a given flight (2h prior).

Audience: Airlines.

Data:

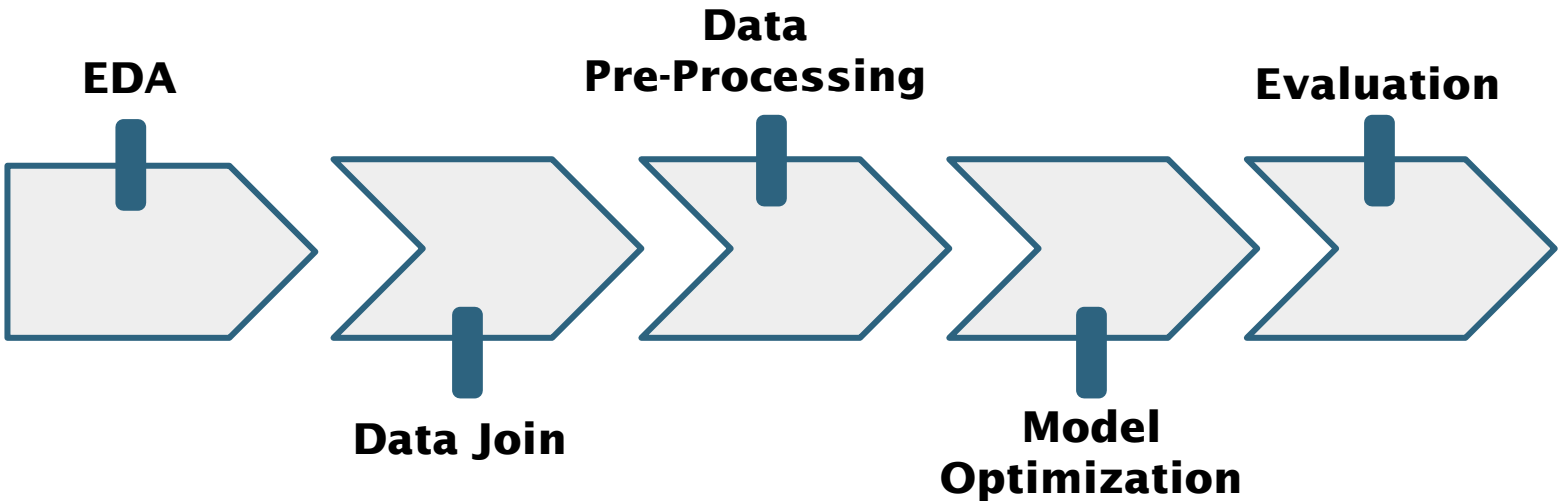
Airlines Data



Weather Station Data

Weather Data

Data Pipeline & Agenda



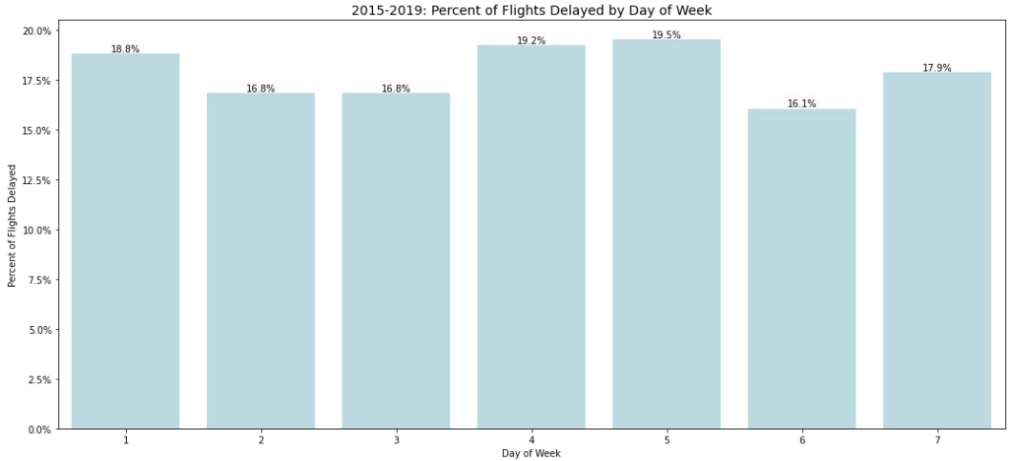
EDA: Airlines Dataset

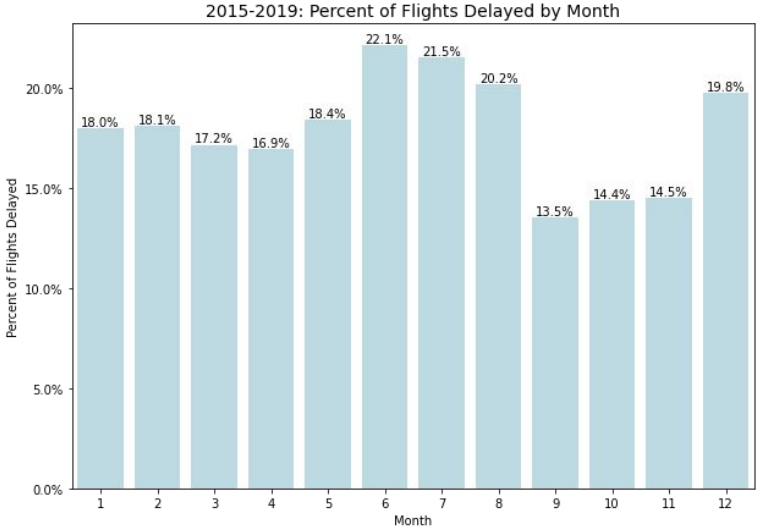
- On-time data for flights from all major US certified air carriers
- Timeframe: 2015-2019
- 63.5 million flights

Departure	# of Flights	% of Flights
On-Time	51,152,008	80.56%
Delayed	11,387,082	17.93%
NA	954,592	1.50%

* EDA performed on the full airlines dataset.

% of Flights Delayed: Day of Week, Month

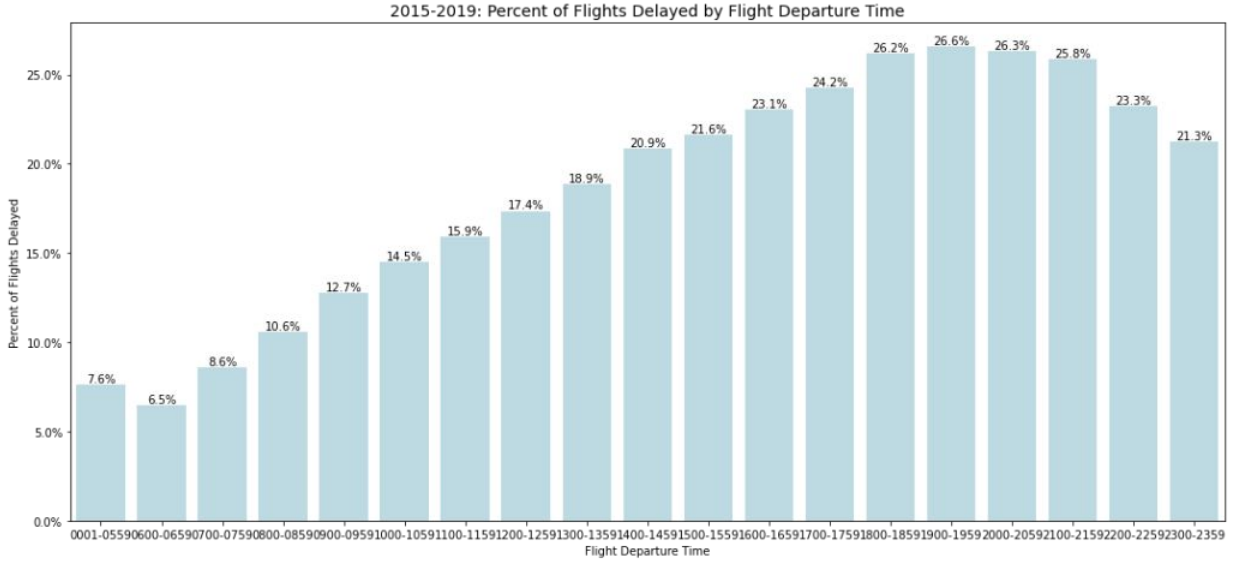




* EDA performed on the full airlines dataset.

Berkeley
UNIVERSITY OF CALIFORNIA

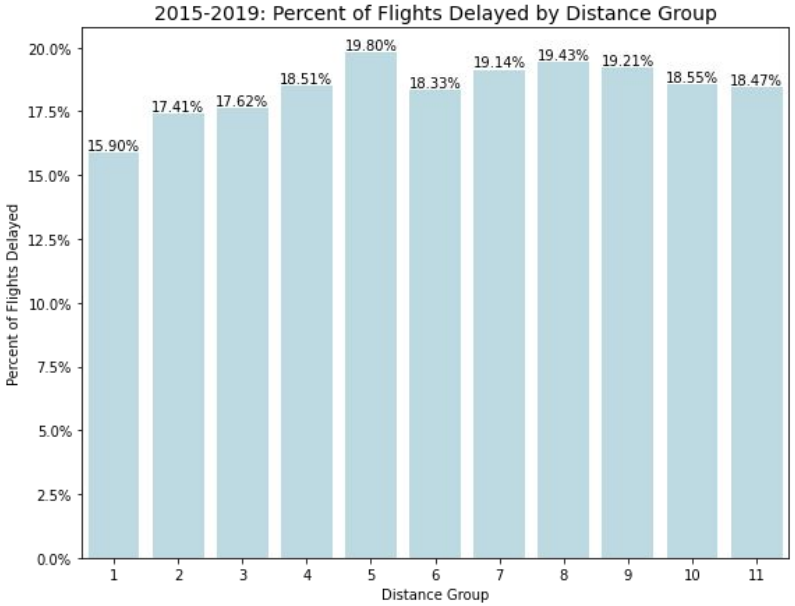
% of Flights Delayed: Time

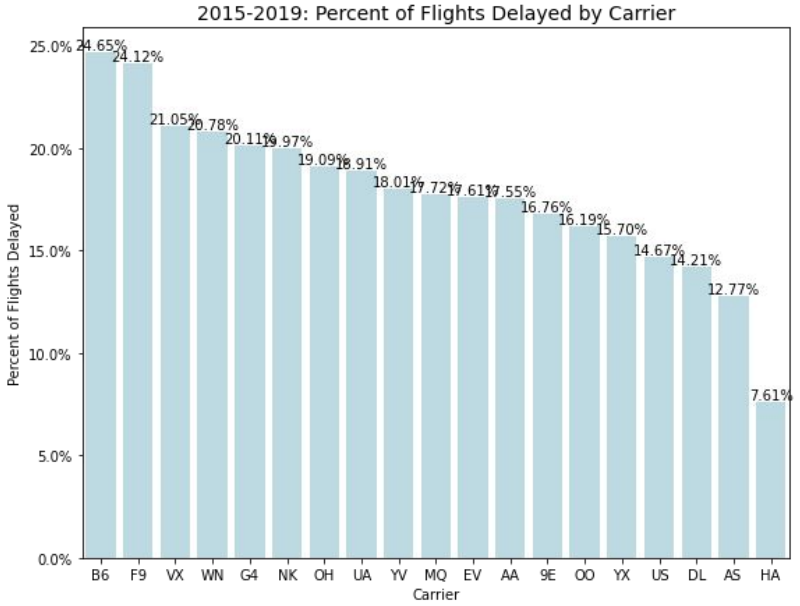


* EDA performed on the full airlines dataset.

Berkeley
UNIVERSITY OF CALIFORNIA

% of Flights Delayed: Distance, Carrier





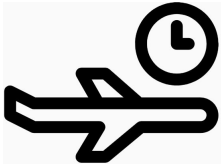
* EDA performed on the full airlines dataset.

Berkeley
UNIVERSITY OF CALIFORNIA

Data Prep, Feature Selection & Engineering

Data Cleaning and Transformations

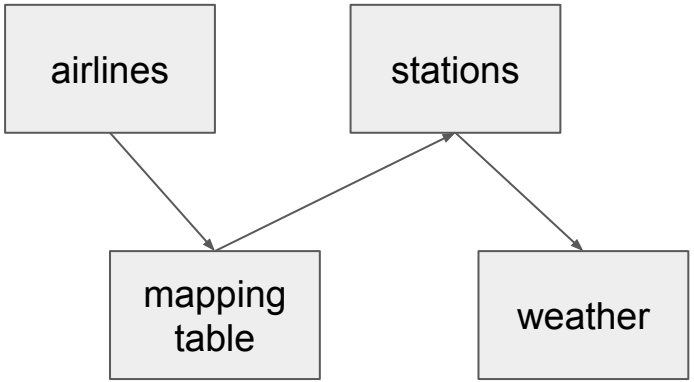
- Weather data parsing and normalization
- Handling of nulls/NaNs
  - Imputation
  - Drop records



Feature Selection & Engineering

- Selection of relevant features
- Derived features:
  - Average flights per airline-airport
  - Average percent delayed flights per airline-airport
  - Average of distance flown per airline-airport
- One-hot Encoded Categorical features

Data Join, Performance & Results

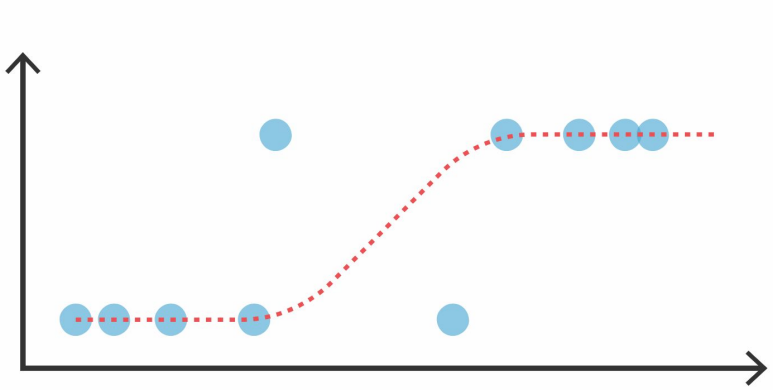


Runtimes	Prep	Join
3-mo Dataset	~3 min	~158 min
Total Dataset	~3 min	~460 min

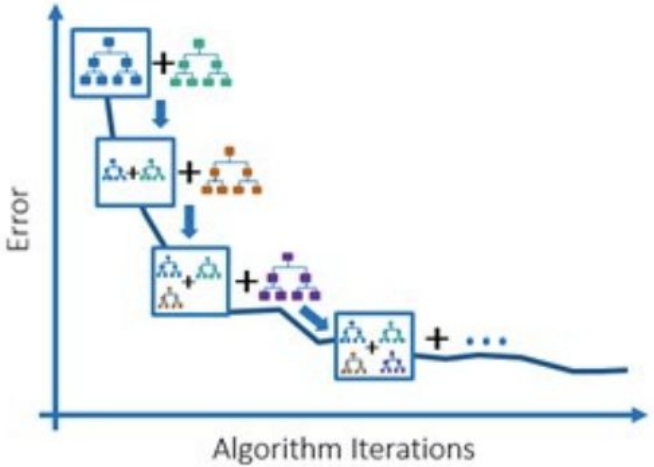
Total Dataset	Training Data	Validation Data	Test Data
# Records	16,831,304	7,078,983	7,268,299

Algorithms

Logistic Regression



Gradient Boosted Trees & XGBoost



Final Model & Advantages

Main Model:

Extreme Gradient Boosting

Further Optimization:

- Gridsearch
- Time Series Cross-Validation

Advantages:

1. Algorithmic optimization
   a. Regularization
   b. Sparsity Awareness
   c. Weighted Quantile Sketch
2. System optimization
   a. Parallelization
   b. Tree Pruning
   c. Hardware Optimization

ML Pipeline Overview

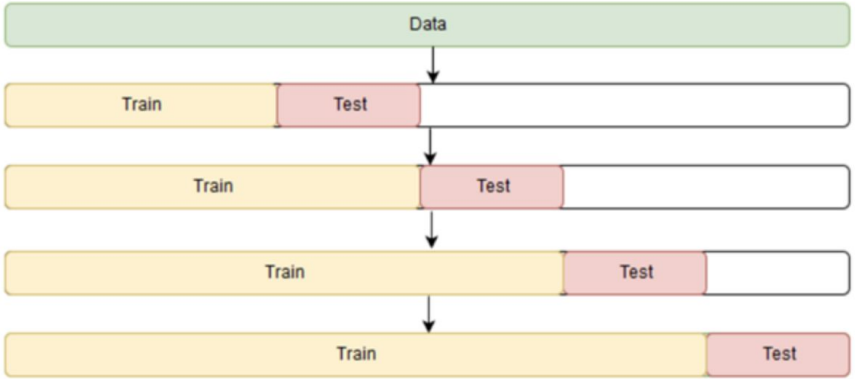
Objectives:

- Address Imbalanced Dataset
  - Down-sampling, Class weights
  - SMOTE
- Custom Pipeline to Input features to fit each models
  - Standardization for numeric features (LR)
  - OHE for categorical features
  - Relevant Feature selection via Chi-Square Selector
- Hyperparameter Tuning
  - Gridsearch
  - K-fold time series split

Cross Validation

Process

1. GridSearch for each model
2. Iterate and choose param set
   - Train & Validate model on each fold
   - Calculate F2 score for each fold
   - Update best validation parameter set
3. Train final model
   - 2015-2018, param from validation
4. Evaluate on final test data



Model Performance & Scalability

	XGBoost		Gradient Boosted Trees		Logistic Regression	
	Train	Test	Train	Test	Train	Test
F1	0.41	0.41	0.42	0.42	0.42	0.42
F2	0.61	0.61	0.57	0.58	0.54	0.54
Recall	**0.90**	**0.90**	0.75	0.76	0.67	0.68
Precision	0.27	0.26	0.29	0.29	0.31	0.30
Runtime (seconds)	1933.2	1462.2	1263.1	1097.4	159.1	604.2

Challenges

- Imbalanced Outcome Variable
- Feature Independence
- Lack of Domain Knowledge
- Multi-Dataset Join
    - Loss of Information

Future Work

- Expand the dataset to more recent years (i.e. 2020)
- Incorporate additional outside datasets to add new features
- Create data pipeline to stream new data in real-time

The End