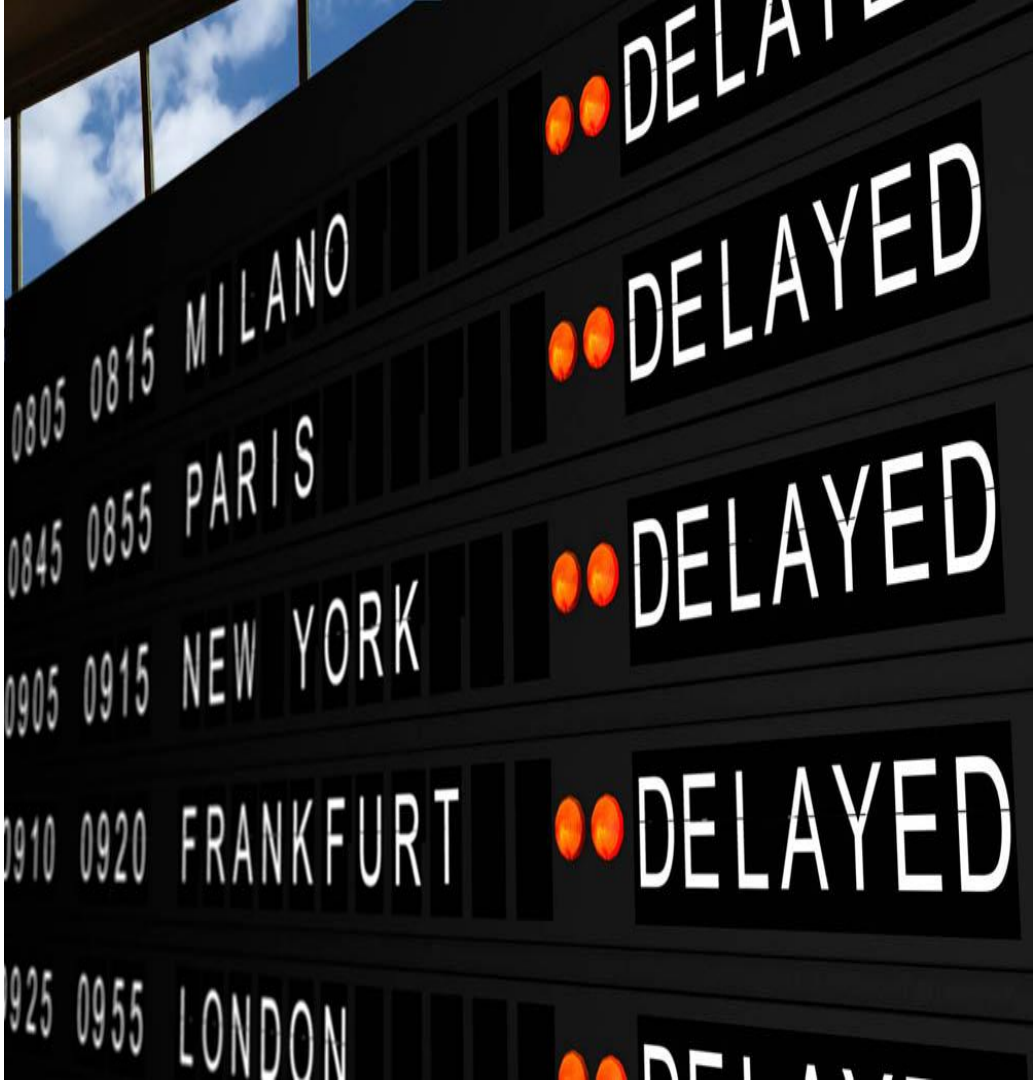
Will your flight be delayed?

Team 12:
Mahesh Arumugam
Stephen Chen
Ramanuj Singh
Jericho Villareal
Radia Wahab



Outline and Contents

- Business Case
- Datasets & Feature Engineering
- EDA
- Algorithms & Evaluation
- Leaderboard
- Novel Ideas
- Scalability Concerns
- Limitations
- Future Scope

Business Case: Flight Delays are Expensive

Context

In aviation, a flight is considered delayed when it leaves 15 minutes past its planned departure time. Flight delays are part of airline proposition to its customers and also a huge factor in measuring an airline's On-Time Performance (OTP) metric, a widely used metric in the aviation industry.

Problem

Flight delays are costly. They are estimated to cost $28 billion annually.*
On average, 1 in 5 flights get delayed daily with delay rates varying across different major airlines. A single flight delay can affect departure times of future flights. This also affects an airline company's airport slots, which is the permission granted by the airport allowing airline companies to schedule a landing or departure during a specific time period.

Solution

Implement an ensemble of machine learning models, at scale, that predict whether a flight would be delayed 2 hours before planned departure time. Public flight records and weather data will be used. Here are the metrics for consideration:

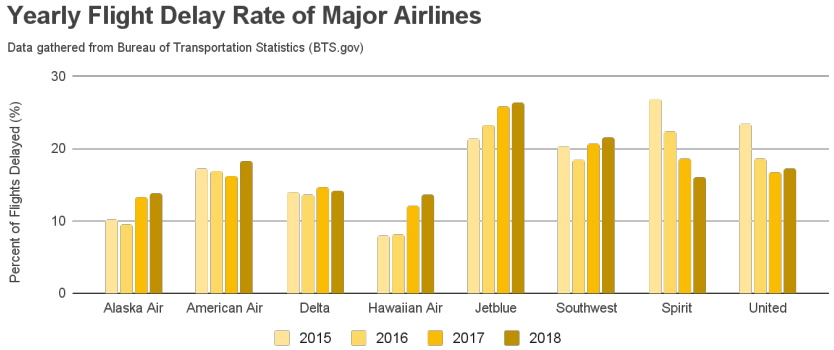
Precision

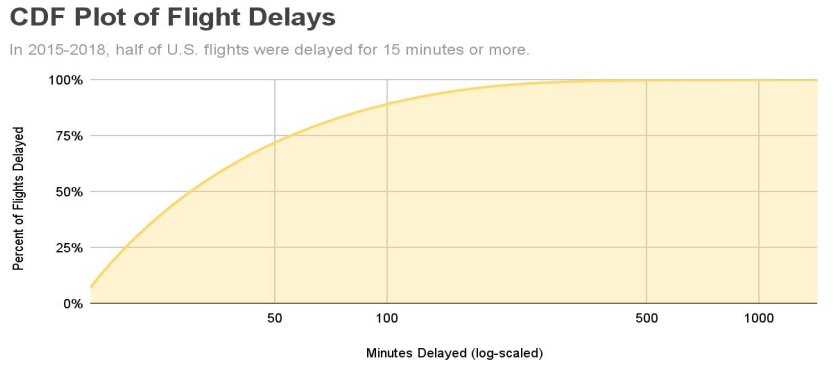
How many delayed flights were successfully predicted in total sample.

Recall

How many delayed flights were successfully predicted in actual delayed flights.

We want to maximize both precision (P) and recall (R) using a metric called Area Under PR. Ideally we want both precision and recall to be as close to 1, however given tradeoffs, it would be best to use a model that maximizes recall.





Datasets

Airlines

Bureau of Transportation Statistics
BTS.gov

31,746,841 rows x 109 columns

Per-flight data of departure time* from 2015 to 2019. Data includes variable which classifies a flight to be delayed or on-time.

* Timezone data included

+

Stations

National Weather Service
NWS.gov

5,004,169 rows x 12 columns

Spatial data on the location of weather stations across multiple geographic regions. Joined with weather dataset to find the closest weather station to an airport.

+

Weather

National Oceanic and Atmospheric Administration (NOAA)

630,904,436 rows x 177 columns

Spatio-temporal weather attributes on across multiple geographic regions. Examples include precipitation, snow depth, wind speed, and air temperature.

Feature Engineering

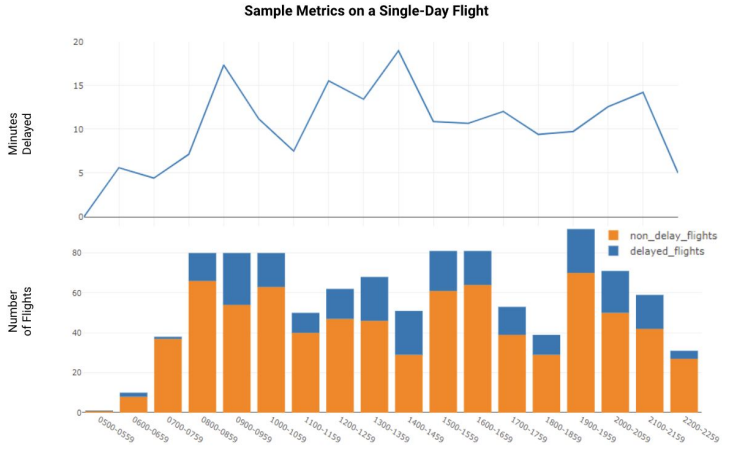
Performance of Airports and Airlines in Each Hour

Aggregate flight on-time performance statistics in 1 hour blocks each day

- Total flights
- Delayed flights
- Percentage of delayed flights
- Average delayed minutes

Metrics for each origin airport

Metrics for each airline at each origin airport



Knock-on Effect (Late Arrival of Incoming Flight)

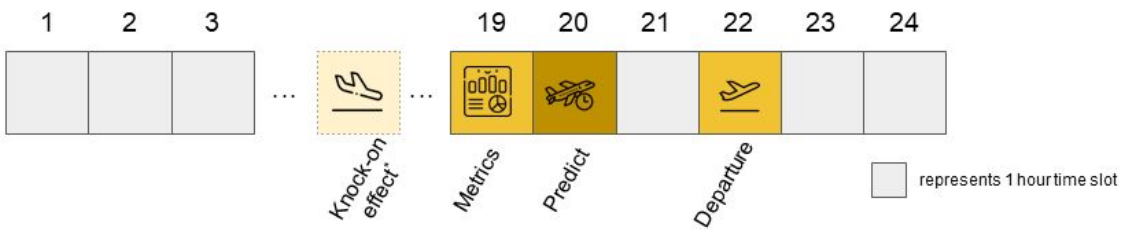
Pagerank of Flight Graph

Feature Engineering

Performance of Airports and Airlines in Each Hour

Knock-on Effect (Late Arrival of Incoming Flight)

- Knock-On effect or delay due to late arrival of an incoming flight
- Looked at arriving flight with the same tail number (TAIL_NUM) that arrived at least 15 minutes late
- Flights with a gap of at-least 2 hours 15 minutes between the time of arrival and the next flight departure



Pagerank of Flight Graph

Feature Engineering

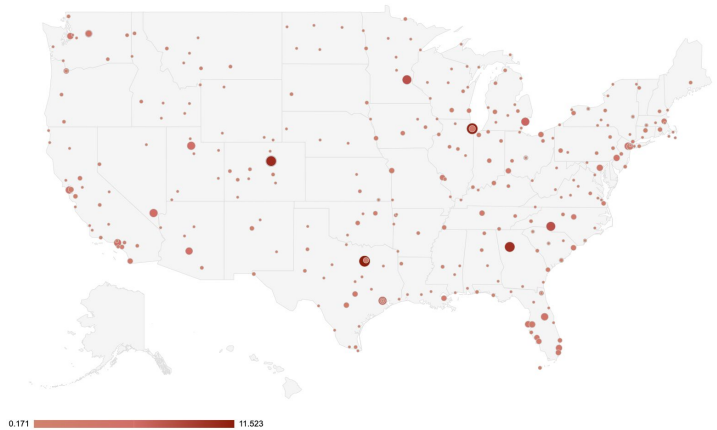
Performance of Airports and Airlines in Each Hour

Knock-on Effect (Late Arrival of Incoming Flight)

Pagerank of Flight Graph

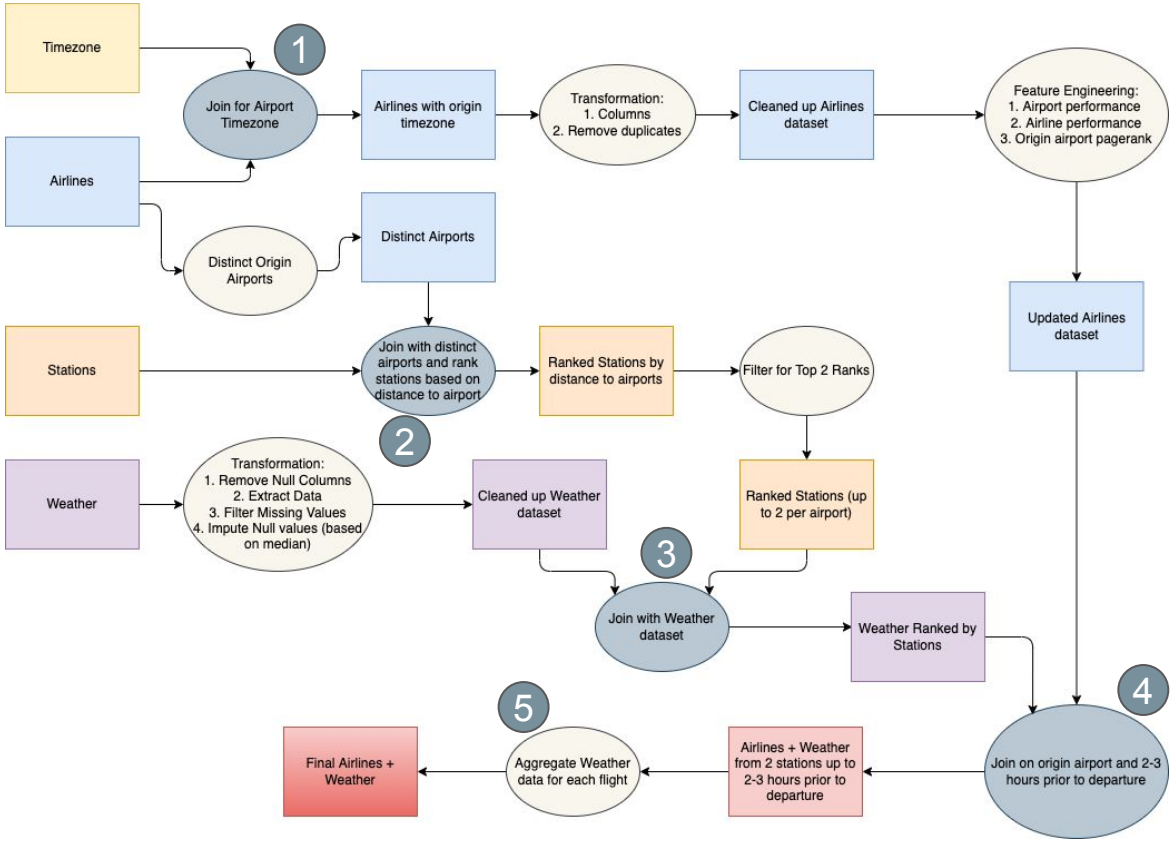
- Create a GraphFrame object with each unique origin and/or destination airport as the vertex.
- Add a directed edge between two airports if there was a connecting flight.
- Use the pagerank method of the GraphFrame library.

Pagerank serves as a proxy for airport movement



Pagerank of each airport in the flight graph represented in the USA map

Full Dataset Join



Time Performance

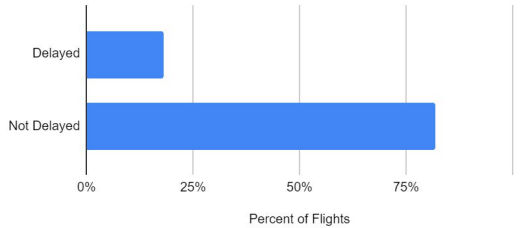
1. 1.97 seconds
2. 3.87 minutes
3. 22.27 minutes
4. 1.88 minutes
5. 2.79 minutes

Exploratory Data Analysis

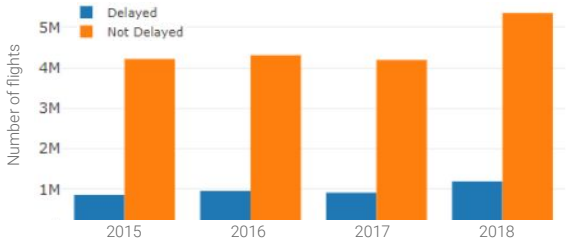
Skewed Class Proportions

Applying classification models on imbalanced data is handled by undersampling (extracting random samples from majority) and upweighting (adding weight to undersampled data).

Imbalanced Data on Flight Delays (1:5 Ratio)



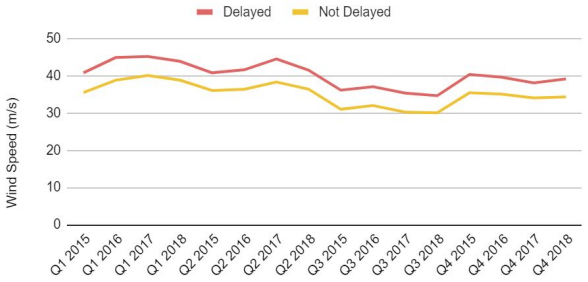
Skewed Data Persisting Over Time

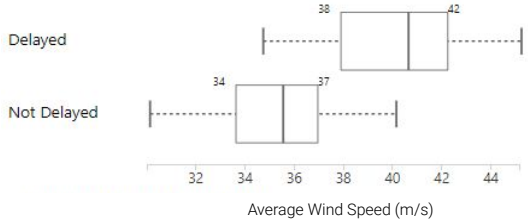


Weather Data Intuition

We are not meteorologists, but have experienced and researched flight delays enough to pull assumptions to initially identify which of the multiple weather variables could affect a flight's departure time.

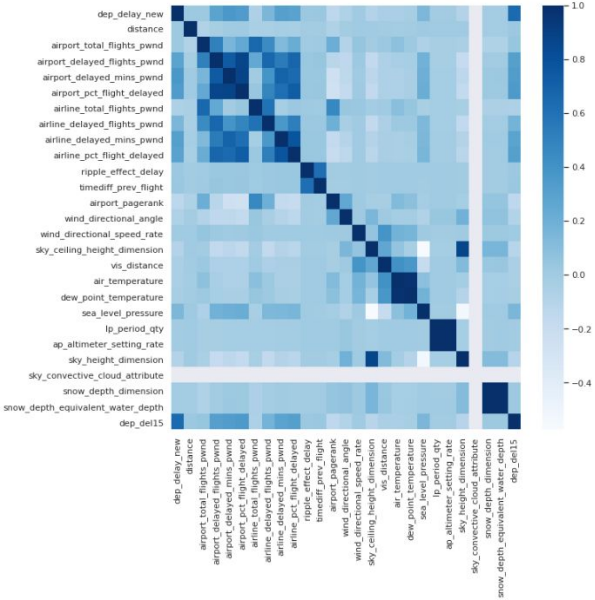
Average Wind Speed per Quarter



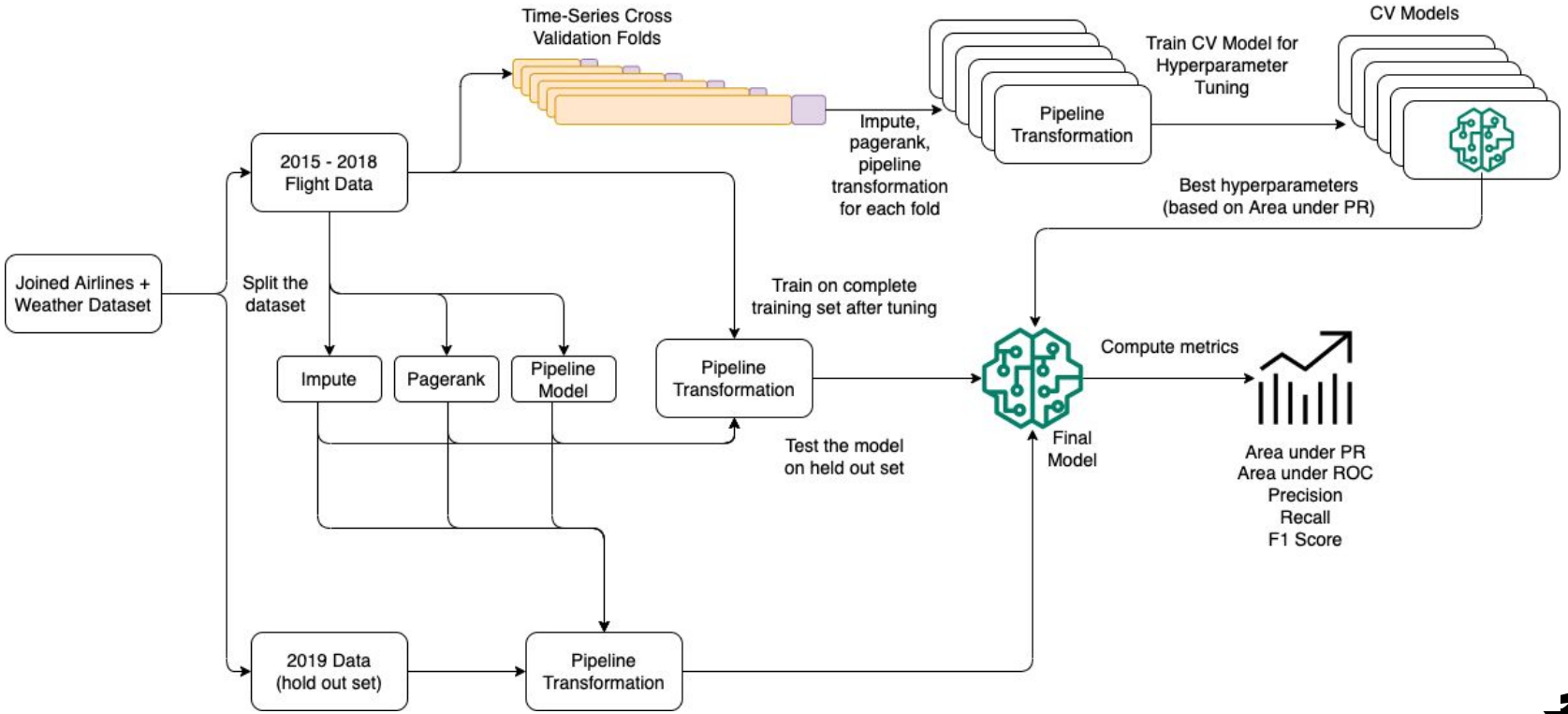


Correlation Matrix

As expected, flight-related metrics are correlated with each other, as well as weather-related data. We are interested in finding out relationships seen in Quadrant 1 and 3 of the matrix.



Machine Learning Pipeline



Algorithms & Evaluation

Logistic Regression

0.362964
Area Under PR

0.290633
Precision

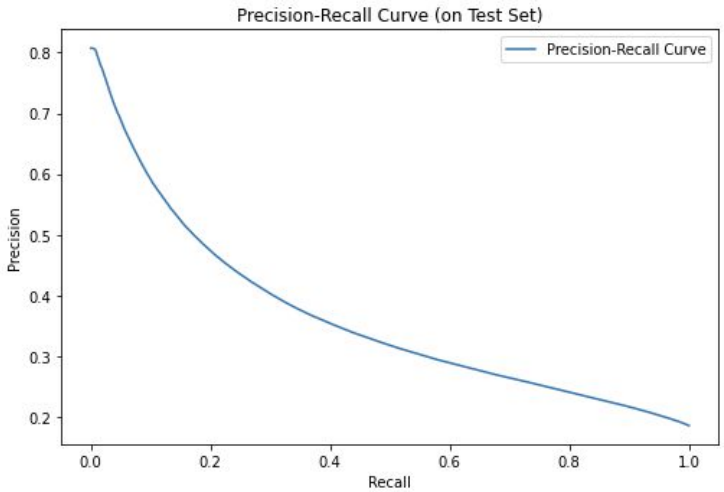
0.591144
Recall

2h26m
CV Time

3m26s
Training Time

3m37s
Evaluation Time

```
fitIntercept:False
regParam: 0.252
maxIter: 52
```



Decision Trees

Gradient Boosted Trees

XGBoost

Algorithms & Evaluation

Logistic Regression

Decision Trees

0.317968
Area Under PR

0.283319
Precision

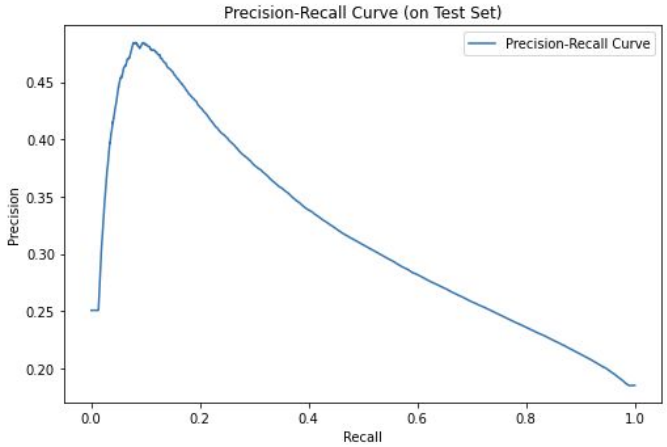
0.592206
Recall

3h19m
CV Time

10m51s
Training Time

2m44s
Evaluation Time

```
maxDepth: 18
maxBins: 12
```



Gradient Boosted Trees

XGBoost

Algorithms & Evaluation

Logistic Regression

Decision Trees

Gradient Boosted Trees

0.346511
Area Under PR

0.295380
Precision

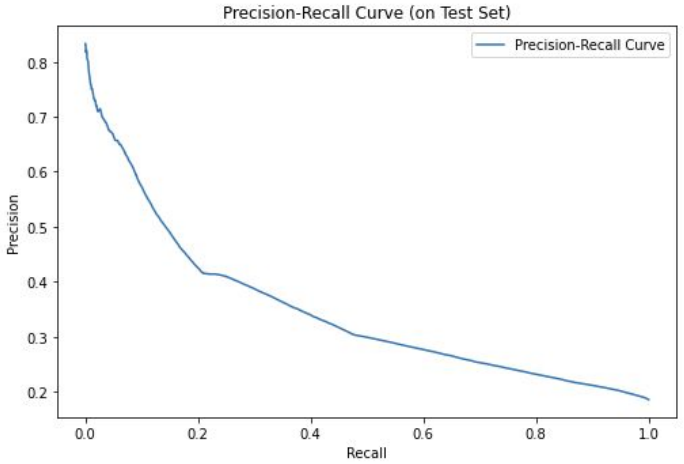
0.515503
Recall

11m57s
CV Time

3m50s
Training Time

2m7s
Evaluation Time

`maxDepth: 2`



XGBoost

Algorithms & Evaluation

Logistic Regression

Decision Trees

Gradient Boosted Trees

XGBoost

0.386705
Area Under PR

0.297037
Precision

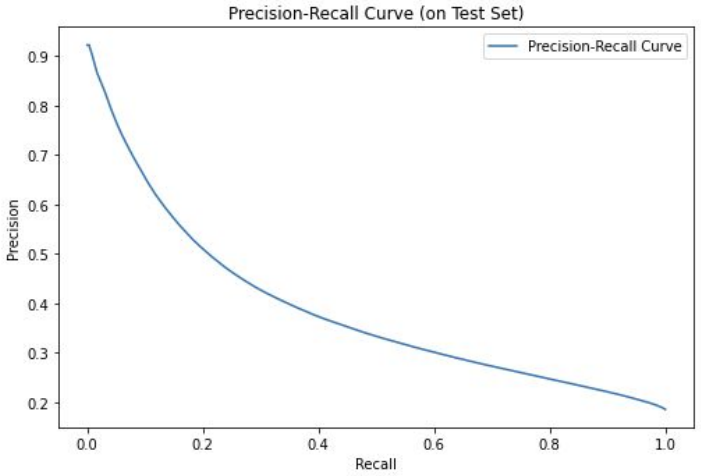
0.612218
Recall

1h7m
CV Time

15m7s
Training Time

5m34s
Evaluation Time

```
max_depth: 6
n_estimators: 32
```



Algorithms & Evaluation

Logistic Regression	Decision Trees	Gradient Boosted Trees	XGBoost
0.362964 Area Under PR	0.317968 Area Under PR	0.346511 Area Under PR	0.386705 Area Under PR
0.290633 Precision	0.283319 Precision	0.295380 Precision	0.297037 Precision
0.591144 Recall	0.592206 Recall	0.515503 Recall	0.612218 Recall

Leaderboard - Gap Analysis

< 5min	**Join Time**	Very fast
15m	**Model Training Time**	Similar training time as most other teams
0.61 0.30	**Recall & Precision Scores**	Similar scores as most other teams on 2019 held-out set
0.387	**Area under PR on Held-out Set (2019)**	Huge **drop-off** from cross-validation to held-out set
58	**Number of Features**	One of the highest
	**Unexplored Algorithms**	Random Forests, CatBoost, Ensemble models, Dimensionality reduction

Novel Ideas

Knock-on Effect (Late Arrival of Incoming Flight)

Pagerank of Flight Graph

Evaluation of 2020 data

4.29M
Airlines Dataset

7.07M
Weather Dataset

4.11M
Joined Dataset

0.195725
Area Under PR

0.160583
Precision

0.386395
Recall

Scalability Challenges & Concerns

Dataset Joins

Problem
Initial attempt to join took about **1.77 days.** Highly inefficient.

Solution
Adjusting the partitions. Final join took < **5 minutes.**

Efficient Memory Usage

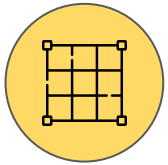
Problem
Improper usage of resources will cause scalability issues.

Solution

Caching

Checkpointing

Un-persisting



Grid vs Random Search

Problem
Grid search for hyperparameter tuning during cross validation took more than **5 hours.**

Solution
Use of Random Search for hyperparameter tuning during cross validation cut down the time by several hours

databricks

Limitations

Trimming Data Set

- Filtering to remove duplicates & cancelled flights

- Undersampling to Balance Data Set

- Only 2 Weather Stations

- For predicting before the two hour window, we only chose to assess one more hour.

Feature Consideration

- Flight delays can also be due to other factors like holiday season etc.

- Widespread effects such as a pandemic, might have effects too.

- Delays can be due to less efficient airport personnel teams.

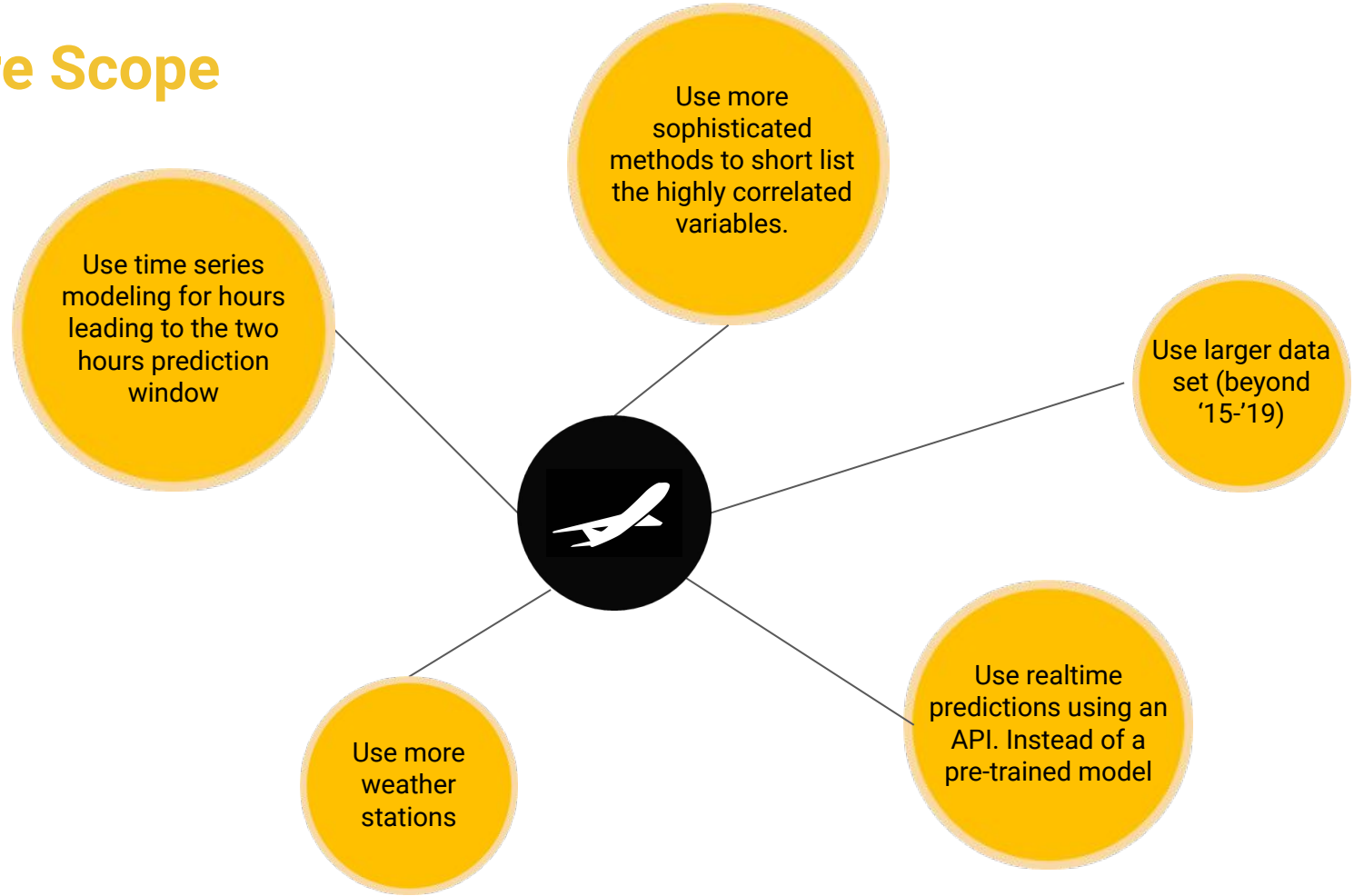
Advanced Model Training

- We only did a binary classification. A regression model to predict how long the delay will be might be valuable too.

- We have 58 features in our model (before encoding and transformation). Dimensionality reduction might be helpful to capture the data with less features.

- And/or, techniques for feature selection might give better results.

Future Scope



Happy Flying