SkyAlliance

Value Proposition:
Predict Flight Delays

Team 4-2

Safiya Alavi, Katya Aukamp, Ainsley Bock, Monica Martin, Clara Rhoades

SkyAlliance

Our Team

Safiya Alavi (Phase Leader)

Katya Aukamp

Monica Martin

Ainsley Bock

Clara Rhoades

What is SkyAlliance ?

SkyAlliance

- Open to all US-based airlines
- Leverages Data Science and Machine Learning
- Provides unique value:
  - customer experience
  - staff satisfaction
  - airport logistics

SkyAlliance

Fast Fact: flight delays cost. a lot.

- Airlines are estimated to lose $7.5-10 billion annually due to flight delays[1]
- In 2019, the total predicted cost experienced by travelers due to flight delays was $2.4 billion[2]

1. U.S. Passenger Carrier Delay Costs | Airlines For America
2. INVESTIGATING THE COSTS AND ECONOMIC IMPACT OF FLIGHT DELAYS IN THE AVIATION INDUSTRY AND THE POTENTIAL STRATEGIES FOR REDUCTION

SkyAlliance

Does delay length matter?

Predicting length of delay enables airlines to tailor logistics & customer service accordingly.

Short Delay (<15 minutes):

- aircraft fuel
- staff scheduled

Medium delay (<60 min):

- ensure lounges are stocked
- extra staff to provide updates
- gate/aircraft switching

Large delay (60+ min):

- extra staff to help reschedule flights
- provide vouchers

WeightedRecall measures the number of true delays predicted. Important to maximize this metric to ensure SkyAlliance has sufficient resources at all times. The cost of missing a real delay is higher than over-predicting one.

SkyAlliance

The problem

Current Stage

We've built baseline models capable of running across all selected algorithms using standardized data inputs. At this point, we have not included feature engineering.

Delay Categories

- No Delay
- Small (0-15 min)
- Medium (15-60 min)
- Large (60+ min)

Why It Matters

Accurately predicting delay severity allows airlines to allocate the right level of support—rebooking staff, lounge access, vouchers—*before* passengers are affected, ultimately building trust and strengthening customer loyalty.

Nosedive: Into the Data

Dataset Quick Facts

ORIGINAL DATASETS

- Reporting Carrier On-Time Performance (OTP)
  - from **Bureau of Transportation Statistics**
- Quality Controlled Local Climatological Data (QCLCD) Publication
  - from **National Oceanic and Atmospheric Administration (NOAA)**
- FAA's Airport Data and Information Portal (FAA-ADIP)
  - from **Federal Aviation Administration**
- Airport Timezones
  - from **Github - Matt Johnson-Pint**
- Aircraft Registration Database (FAA-ARD)
  - from **Federal Aviation Administration**

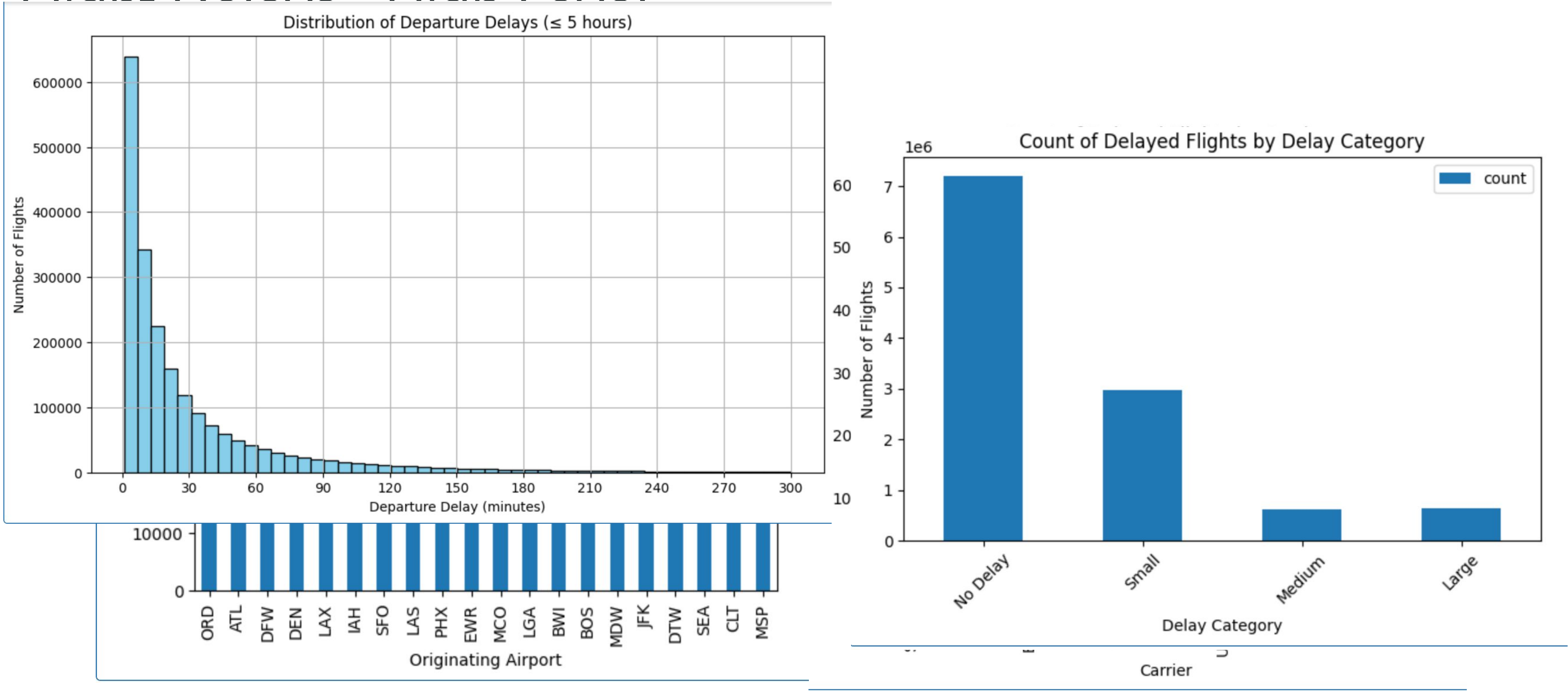
JOINED DATASET

- Did a custom join with the OTP data on the QCLCD, FAA-ADIP, and FAA-ARD

PROCESSED DATA OVERVIEW

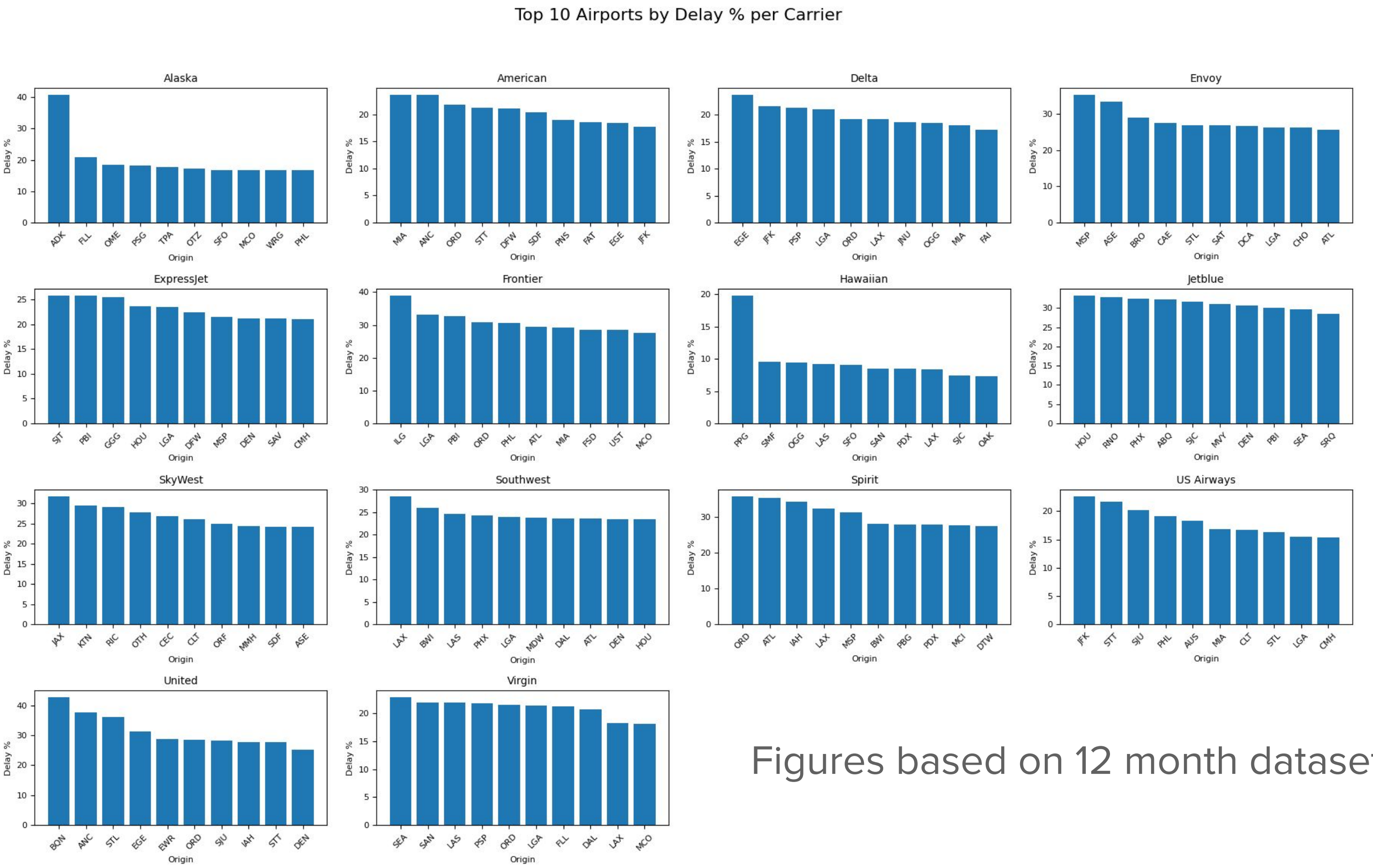
- 23 features
- training samples
  - Jan - Sep 2015
  - 2,839,442 records (~67%)
- testing samples:
  - Oct - Dec 2015
  - 1,775,633 records (~33%)

Flight Delays - High Level



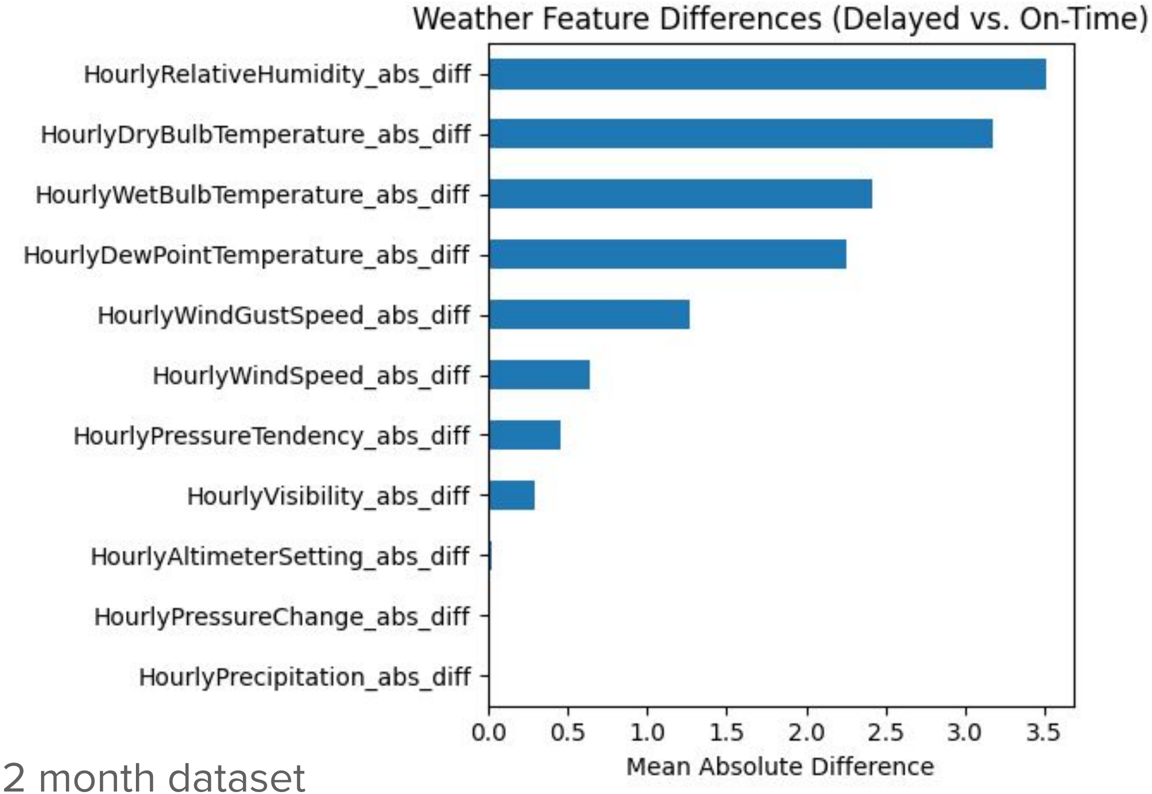
Figures based on 12 month dataset

Delay by Percentage



Figures based on 12 month dataset

Weather Component



Figures based on 12 month dataset

Summary of Our Data

It is skewed

We see that delays are not evenly distributed by location/carrier and time period and so to model this data we will need to rebalance this data

Weather

There seem to be no super strong indicators in weather besides humidity and temperature

Cleaning

- Virgin Atlantic is no longer a US carrier is now mapped to Alaska
- US Airlines merged with American airlines and so for current day predictions will be mapped under American

Data Preprocessing

- Dropped duplicates from the dataset.
- Excluded canceled flights from the dataset ➔ the focus of this project is to enable SkyAlliance to allocate resources efficiently in response to delays, not cancellations.
    - Cancellations often trigger a different set of responses, such as rerouting and refunds, that fall outside the scope of proactive delay management.
- In the columns relating to the departure being delayed, if a row was Null, it was dropped.
    - Dropped rows with Nulls in any of the selected columns relating to weather or airport information
    - Automatically dropped columns with over 70% Nulls
    - Dropped rows from airports that are currently closed

Data Preprocessing

- Merged airline carriers that were acquired by others
  - Including Virgin Atlantic and US Airlines
- UTC timestamp conversions for departure and arrival times
- Created a column to account for flights which use the same aircraft as a previous flight that was delayed
- Balanced the data by downsampling the No Delay category to the amount of data in the delay categories combined, then upsampled each individual delay type
  - Even distribution of about 25% per category
- Data splitting
  - Train: 1 Jan 2015 - 31 Aug 2015
  - Test: 1 Sep 2015 - 31 Dec 2015
    - Train-Test Split: 67%/33%

More on Null Handling In Weather and Runway Data

Nulls & Data Cleaning in Weather Data

The weather data was also cleaned after the join to the flight data

Feature	Null Count	Dropped/Kept	Note
station	0	N/A	
date	0	N/A	
HourlyVisibility	32,400	Dropped	Smaller Stations don't always report
HourlyDewPointTemperature	33,161	Dropped	Smaller Stations don't always report
HourlyDryBulbTemperature	26,301	Dropped	Smaller Stations don't always report
HourlyWetBulbTemperature	121,017	Dropped	Smaller Stations don't always report
HourlyRelativeHumidity	33,469	Dropped	Smaller Stations don't always report
HourlyWindSpeed,	27,742	Dropped	Smaller Stations don't always report

Nulls & Data Cleaning in Runway Data

The features that are numeric they were all used for averages so the nulls didn't factor into the calculation

Feature	Null Count	Dropped/Kept	Note
Site_Id	0		
Loc_Id	0		
Runway_Id	0		
Length	0		
Width	0		
Base_Obstacle_Clearance_Slope	9009	Dropped	Not all runways have obstacles
Base LDA	14,983	Dropped	Some smaller airports don't report to this level
Base TORA	14,981	Dropped	Some smaller airports don't report to this level

Current Data Features

Time (OHE)

- Year
- Month
- Day of Month
- CRS Elapsed Time (Num)
- Origin Timezone
- Dest. Timezone

Hourly Weather (Num)

- Dew Point Temperature
- Dry Bulb Temperature
- Relative Humidity
- Visibility
- Wet Bulb Temperature
- Wind Speed
- Elevation

Flights (OHE)

- Carrier Airline ID
- Tail Num
- Origin Airport ID
- Dest. Airport ID

OHE: One Hot Encoding
Num: Numeric

Current Data Features continued

FAA Airport Data (Num)

- Site ID
- Location ID
- Longitude
- Latitude

FAA Runway Data (Num)

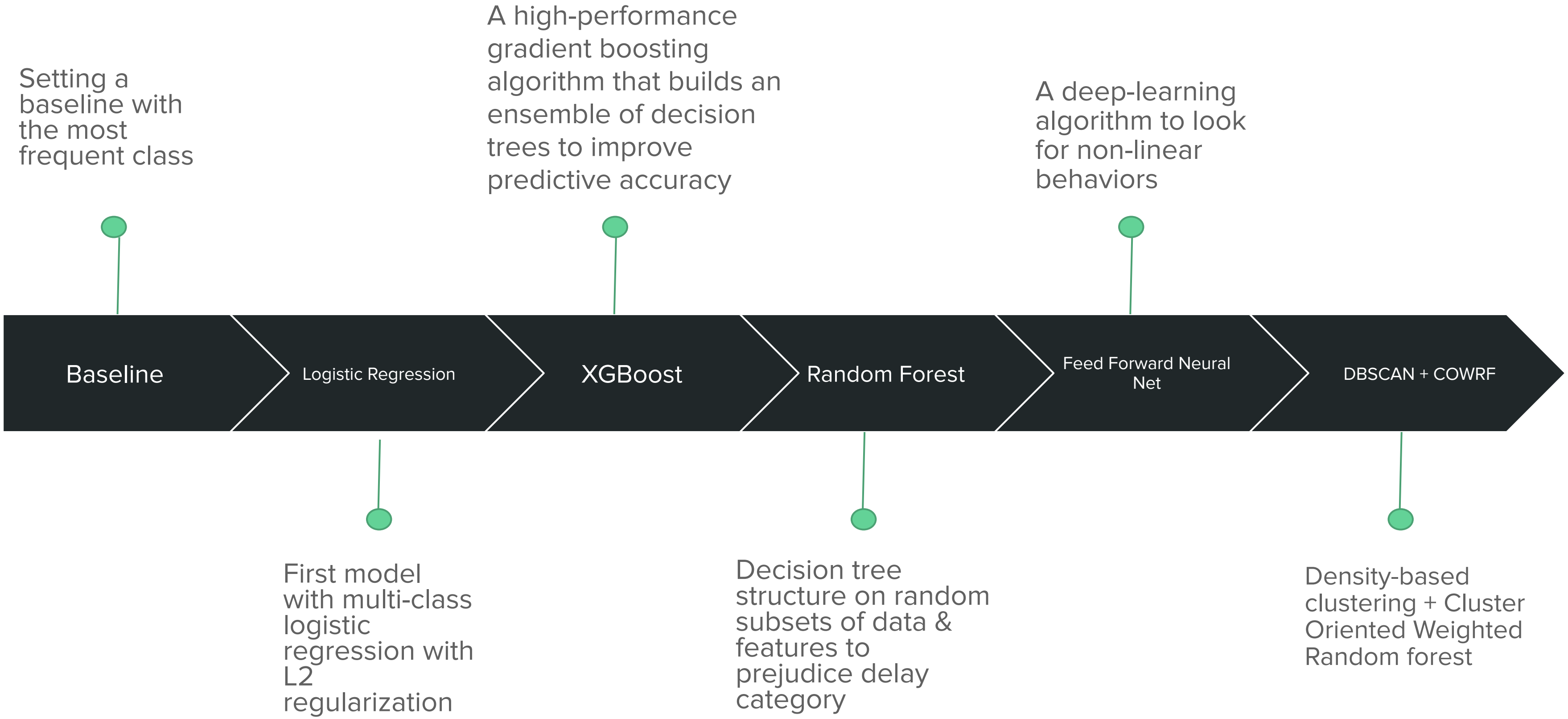
- Site ID
- Location ID
- Runway ID
- Length
- Width
- Base Obstacle Clearance Slope
- Base Landing Distance Available
- Take Off Runway Available

Engineered Features (Bool and Num)

- Prior flight existing
- If the prior flight was delayed
- The average of the base obstacle slope
- Average length of the runway
- Average width of the runway
- Average of the landing distance available
- Average take off distance available
- Number of runways at each airport

OHE: One Hot Encoding
Num: Numeric

Algorithms



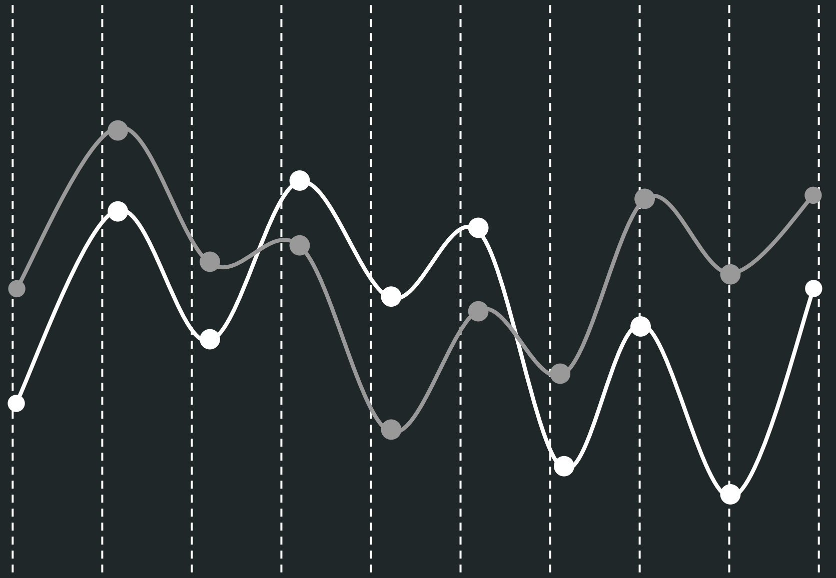
All models are evaluated using cross validation (specific to time series data) split from the 9 month training data.

Evaluation Metric

WeightedRecall is the metric we will train our models to maximize on.

- Recall is prioritized over precision because the cost of missing a real delay (false negative) is higher than over-predicting one (false positive).
  - Missed delay ➜ insufficient staffing, customer dissatisfaction, and operational disruption.
  - False alarm ➜ some over-preparation but helps ensure readiness and prevents service breakdowns.
- Maximizing recall supports SkyAlliance’s ability to deliver proactive, cost-effective responses, ultimately improving both customer experience and operational resilience.

Initial Outcomes



Baseline Model

Our baseline model always predicts the most frequent class: no delay

Our baseline model's unweighted recall is 0.25 because it has perfect recall for `no delay` and 0 recall for small, medium, and big delay.

Averaged over each class, this leads to a recall of 0.25.

```
model_name: baseline_model

recall: 0.25

precision: 0.15027762784747467

f1: 0.45135403750457836

accuracy: 0.6011105113898987
```

Logistic Regression - Test Dataset Findings

Model	Notes	Unweighted Recall	Unweighted Precision	F1 Score	Accuracy
Baseline		0.25	0.17	0.54	0.67
Vanilla Logistic Regression		0.258	0.33	0.55	0.67
L2 Log Regression	lambda = 1.0 100 epochs	0.25	0.23	0.54	0.67
L1 Log Regression	lambda = 1.0 100 epochs	0.25	0.17	0.54	0.67

Per Class Metrics for Vanilla Logistic Regression:

	Class	Precision	Recall	F1 Score
0	no delay	0.677310	0.980953	0.801332
1	small delay	0.324004	0.044811	0.078733
2	medium delay	0.222863	0.004322	0.008479
3	large delay	0.090487	0.000474	0.000943

XGBoost Classifier

Background:

XGBoost can naturally capture complex nonlinear relationships and feature interactions, leading to better performance on real-world data.

Initial Model Parameter Tuning:

Below is a table of the **average** WeightedRecall **across 5 cross validation folds** with 2 different learning rate and tree depths tested:

Learning Rate → **Depth ↓**	**0.05**	**0.10**
8	0.3063	**0.3083**
10	0.3061	0.3079

Best hyperparameters from CV leading to highest WeightedRecall: [List out the hyperparameters used for final model]

- Learning Rate: 0.1
- Depth: 8

XGBoost Classifier

Evaluation on the Training Set:

- **Yields Weighted Recall of [4 sig figs]**
- **Yields Weighted Precision of [4 sig figs]**
- **Yields Weighted F1 Score of [4 sig figs]**

[Description of results on each delay category]

Label	Recall	Precision	F1
No Delay			
Small Delay			
Medium Delay			
Large Delay			

Evaluation on the Testing Set:

- **Yields Weighted Recall of 0.5333**
- **Yields Weighted Precision of 0.7672**
- **Yields Weighted F1 Score of 0.6191**

[Description of results on each delay category]

Label	Recall	Precision	F1
No Delay			
Small Delay			
Medium Delay			
Large Delay			

Random Forest Classifier

Key Insights (Generally):

- Reduces Overfitting
- Performance is consistent between Train and Test
- Insights into feature Importance

Performance Baseline Results:

- Biased towards “No-Delay” class
- Performance is consistent between Train and Test
- Low recall scores for the delay categories

Train:

* Default Parameters

	label	1.2 recall	1.2 precision	1.2 f1
1	no delay	0.7990844983298281	0.28744375763453667	0.4227995905427566
2	small delay	0.24816283579882945	0.2762805648475812	0.261467942447823
3	large delay	0.0118674654648052...	0.3434972822484423	0.0229422989228364...
4	medium delay	0.0475776037133739...	0.41391494552643754	0.0853451752655111

Test:

* Default Parameters

	label	1.2 recall	1.2 precision	1.2 f1
1	no delay	0.8012540099154273	0.23636018223463245	0.36503846079324176
2	small delay	0.27575938058368077	0.2612696690179056	0.26831904958263214
3	large delay	0.0131453397486651...	0.32993630573248406	0.0252833393531760...
4	medium delay	0.0182935043241641...	0.4615907545887152	0.03519228775785218

Feed Forward Neural Net

- Only guesses the majority class
  - Fails to outperform baseline
  - Attempted Input Layers: [5610, 32, 8, 4], [5610, 256, 32, 4], [5610, 16, 8, 4]
  - Long training times

Train

	Class	Precision	Recall	F1 Score
0	no delay	0.609765	1.0	0.757583
1	small delay	0.000000	0.0	0.000000
2	medium delay	0.000000	0.0	0.000000
3	large delay	0.000000	0.0	0.000000

Test

	Class	Precision	Recall	F1 Score
0	no delay	0.671471	1.0	0.803449
1	small delay	0.000000	0.0	0.000000
2	medium delay	0.000000	0.0	0.000000
3	large delay	0.000000	0.0	0.000000

DBSCAN + COWRF

Create unsupervised clusters of flights

Add the clusters to the data as additional data

Run Random Forest per cluster

	Precision	Recall	F1 Score
No Delay	0.671485	1.0	0.803459
Small Delay	0.714286	0.000398	0.000795
Medium Delay	0.628571	0.000235	0.000469
Big Delay	0.0	0.0	0.00

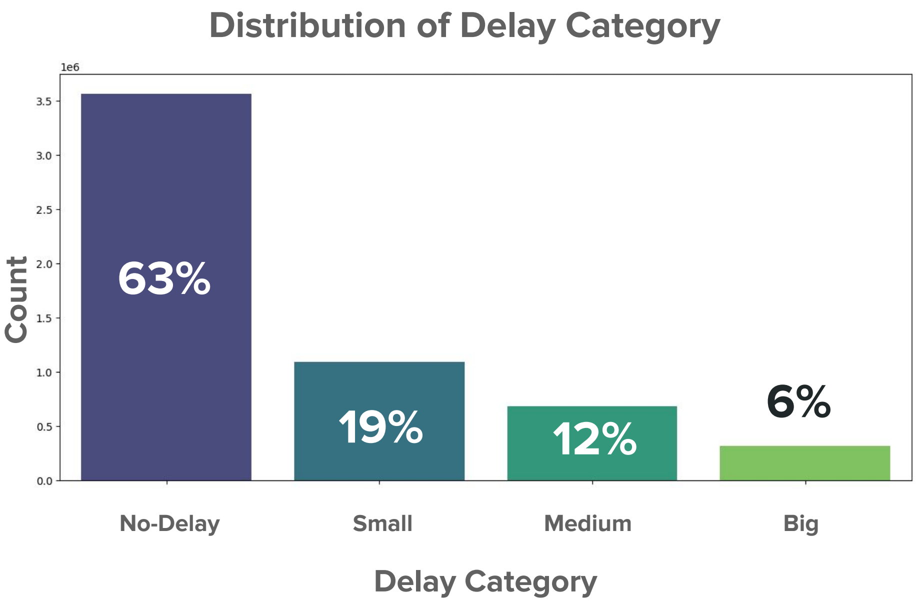
References

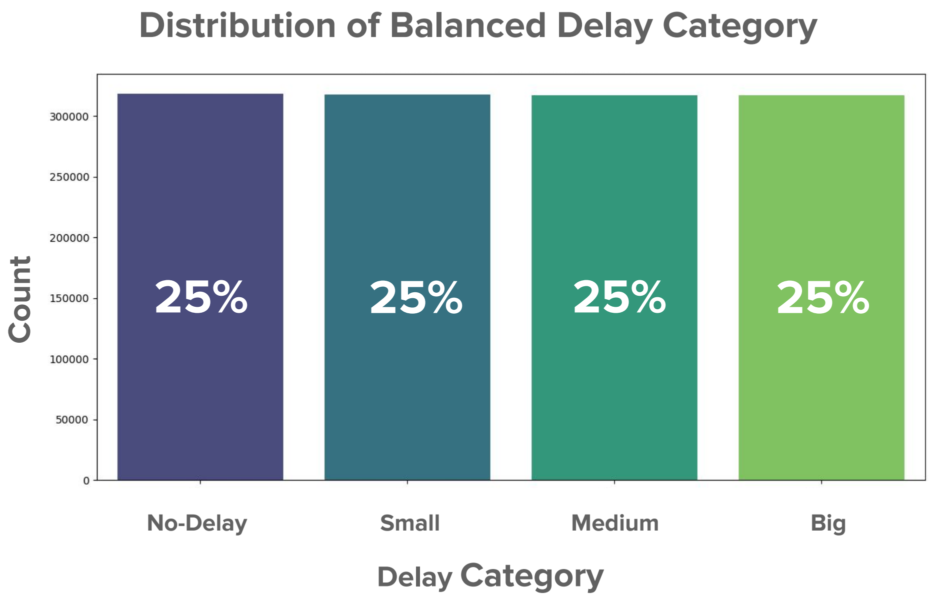
- ChatGPT: SkyAlliance logos, quirky section titles, light editing
- US Passenger Carrier Delay Costs:
- INVESTIGATING THE COSTS AND ECONOMIC IMPACT OF FLIGHT INVESTIGATING THE COSTS AND ECONOMIC IMPACT OF FLIGHT DELAYS IN THE AVIATION INDUSTRY AND THE POTENTIAL DELAYS IN THE AVIATION INDUSTRY AND THE POTENTIAL STRATEGIES FOR REDUCTION STRATEGIES FOR REDUCTION
- BTS TranStats: Airline On-Time Statistics and Delay Causes
- Investigating the Costs and Economic Impact of Flight Delays

Backup Slides

Data Balance Experiments:

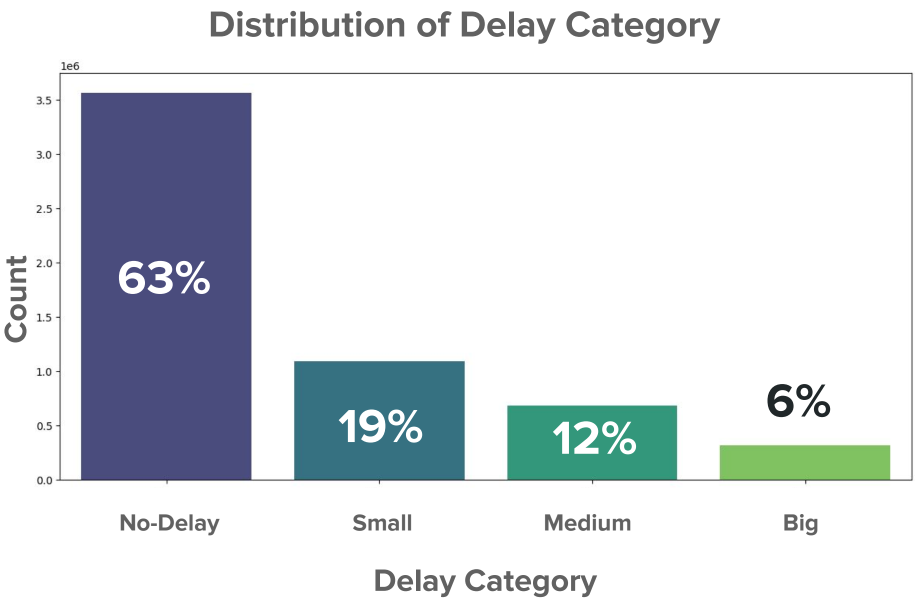
Method 1: DownSample **all classes** to **minority** class size

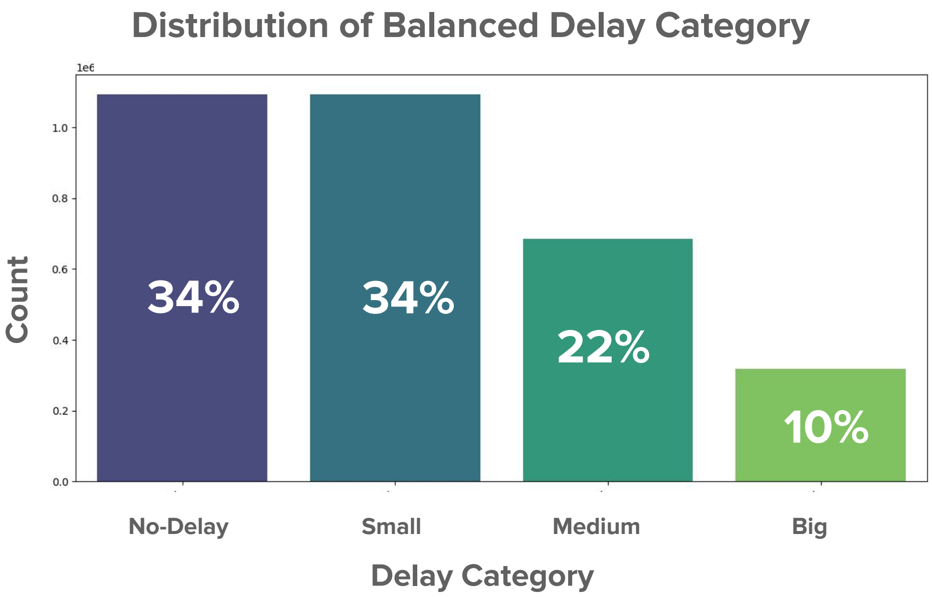




Data Balance Experiments:

Method 2: Down Sample **majority only** class to the **2nd largest** class size





Model Name

Background:
Opt to include one sentence background on the model

Initial Model Parameter Tuning:
Either fill in the table with the WeightedRecall from either tuning combination or list the different parameters tested

Learning Rate → **Depth ↓**	**0.05**	**0.10**
4		
6		

Best hyperparameters from CV leading to highest WeightedRecall: [List out the hyperparameters used for final model]

- Learning Rate:
- Depth:

Model Name

Evaluation on the Training Set:

- **Yields Weighted Recall of [4 sig figs]**
- **Yields Weighted Precision of [4 sig figs]**
- **Yields Weighted F1 Score of [4 sig figs]**

[Description of results on each delay category]

Label	Recall	Precision	F1
No Delay			
Small Delay			
Medium Delay			
Large Delay			

Evaluation on the Testing Set:

- **Yields Weighted Recall of [4 sig figs]**
- **Yields Weighted Precision of [4 sig figs]**
- **Yields Weighted F1 Score of [4 sig figs]**

[Description of results on each delay category]

Label	Recall	Precision	F1
No Delay			
Small Delay			
Medium Delay			
Large Delay			