U.S. Airline Labor Shortage Impact Reduction: Flight Delay Prediction

W261 - Final Project: Section 3 Group 2

Shehzad Shahbuddin

Mohamed Sondo

Max Hoff

Charles Carlson

Justin Wong

Outline

Abstract / Project Description

Problem Statement

- Recent U.S. labor shortages have caused flight cancellations
- ***Goal:*** Predict cancellation of flights using weather data in order to help re-allocate workers to higher probability flights

Focus Areas

- Feature engineering
  - Net flow
  - Weather Prediction
  - Cascading Delays
- Modeling approaches
  - Random Forest
  - XGBoost
  - Neural Nets

Challenges

- Ideation on feature engineering
- Selecting focus areas
- Scalability
- Domain Knowledge
- Time

Results

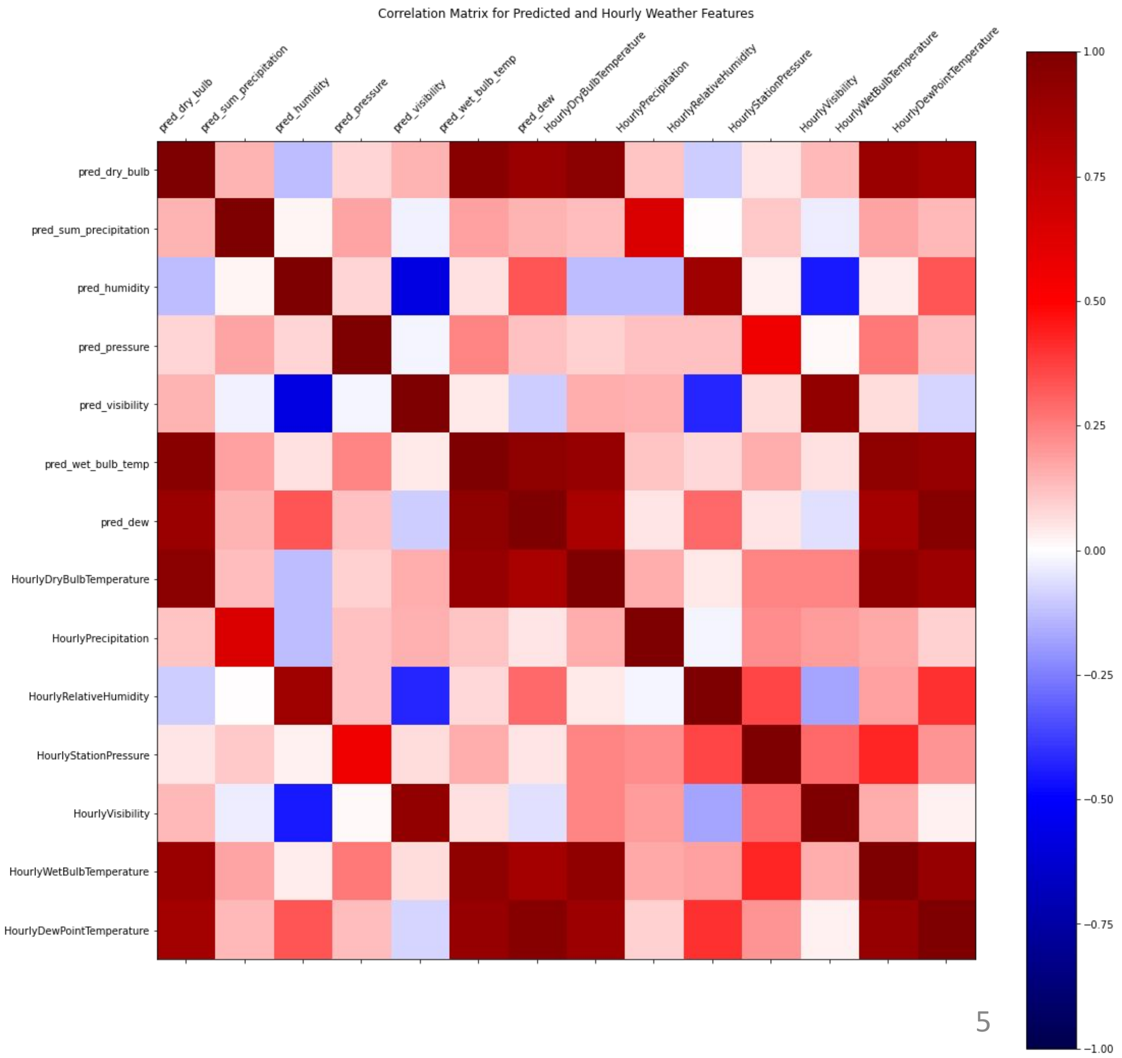
- Model: Random Forest
- Weighted Precision: 0.764
- Weighted Recall: 0.806
- Weighted F1: 0.784

Feature Engineering

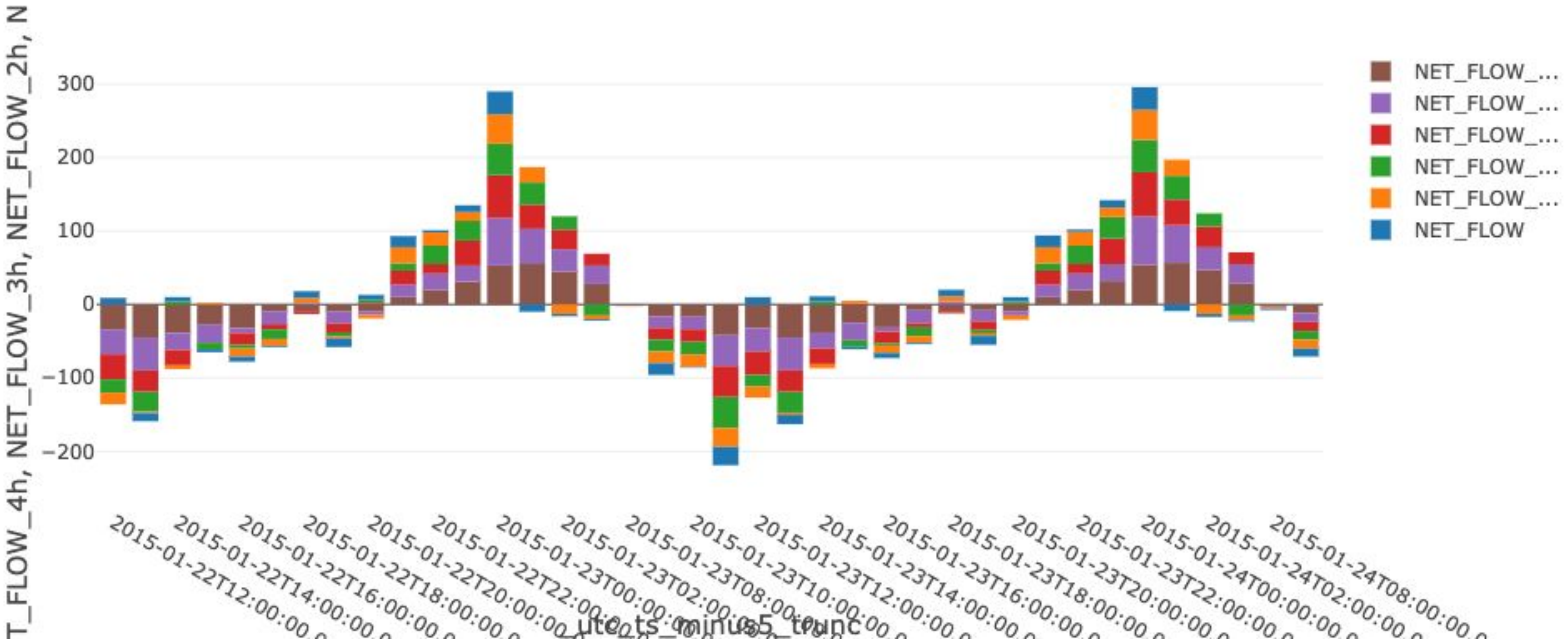
Predicted Weather

Linear Regression

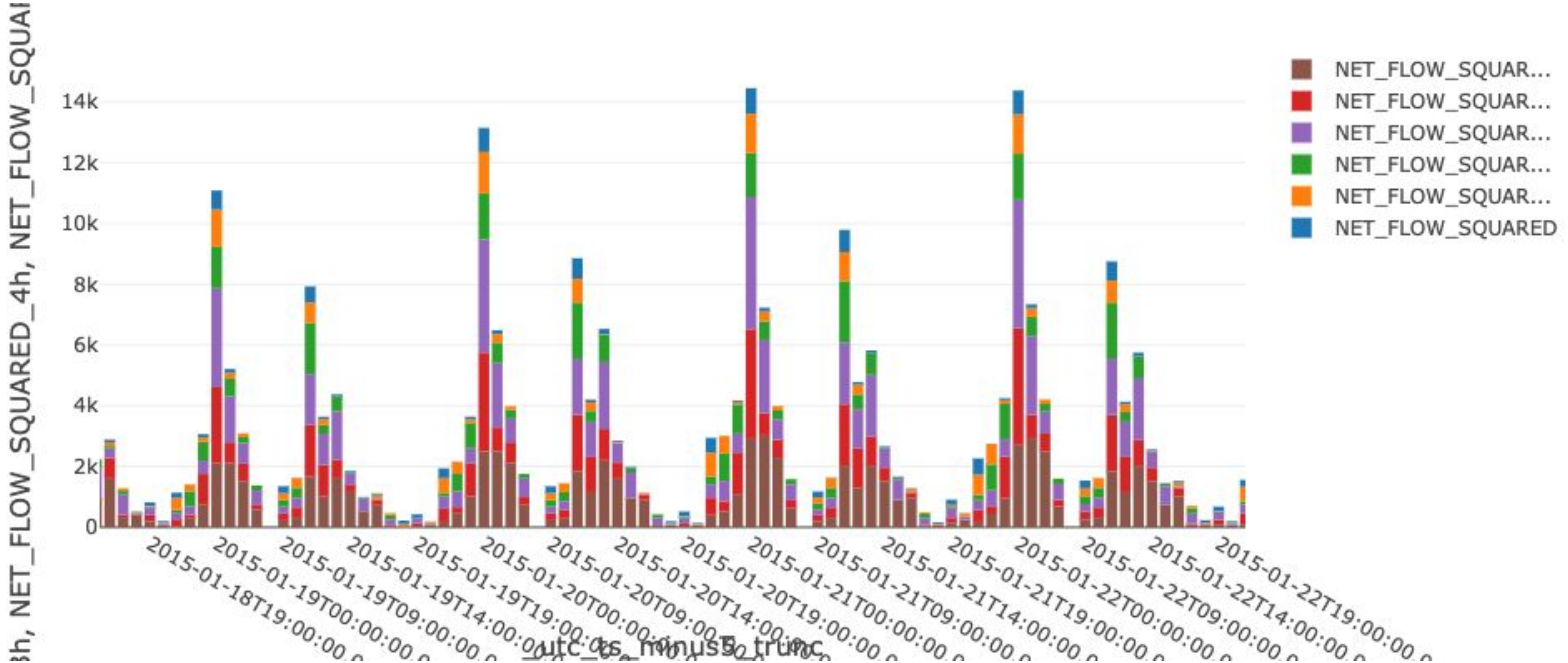
- Outcome variables being basic meteorological variables like temperature.
- Regressors include the same basic meteorological variables from 2 hours before the flight; aggregates of these variables for the hours preceding; the latitude, longitude, and elevation; and the day of the year (doy) and the doy squared.
- The average R-Squared was 0.815 with the highest R-Squared being 0.985.



Windowed Net Flow



Windowed Squared Net Flow



Page Rank - Graph

	ORIGIN	collect_set(DEST)
1	ABE	["SRQ", "SAV", "PGD", "PHL", "ORD", "CLT", "DTW", "FLL", "MDW", "PIE", "SFB", "ATL", "MYR", "BNA", "MDT", "EWR"]
2	ABI	["IAH", "DFW", "GRK"]
3	ABQ	["SEA", "SNA", "DFW", "ORD", "MSP", "IAH", "HOU", "DEN", "LAX", "ATL", "LAS", "MCI", "PHX", "SLC", "OAK", "CLT", "SAF", "AUS", "DAL", "PDX", "MDW", "SFO", "MCO", "SJC", "SFB", "JFK", "SAN", "SAT", "BWI"]
4	ABR	["MSP"]
5	ABY	["ATL"]
6	ACK	["HPN", "BOS", "JFK", "DCA", "PHL", "CLT", "ORD", "LGA", "EWR"]
7	ACT	["DFW", "DEN"]
8	ACV	["SFO", "DEN", "LAX", "PHX"]
9	ACY	["FLL", "DTW", "MIA", "RSW", "MSY", "PBI", "MCO", "TPA", "ATL", "MYR", "BOS", "ORD", "SJU"]
10	ADK	["ANC", "CDB"]

Showing all 381 rows.

Page Rank - Scores

Top 10 Airports by Page Rank Scores
1. DEN
2. ORD
3. DFW
4. ATL
5. CLT
6. MSP
7. IAH
8. DTW
9. LAX
10. LAS

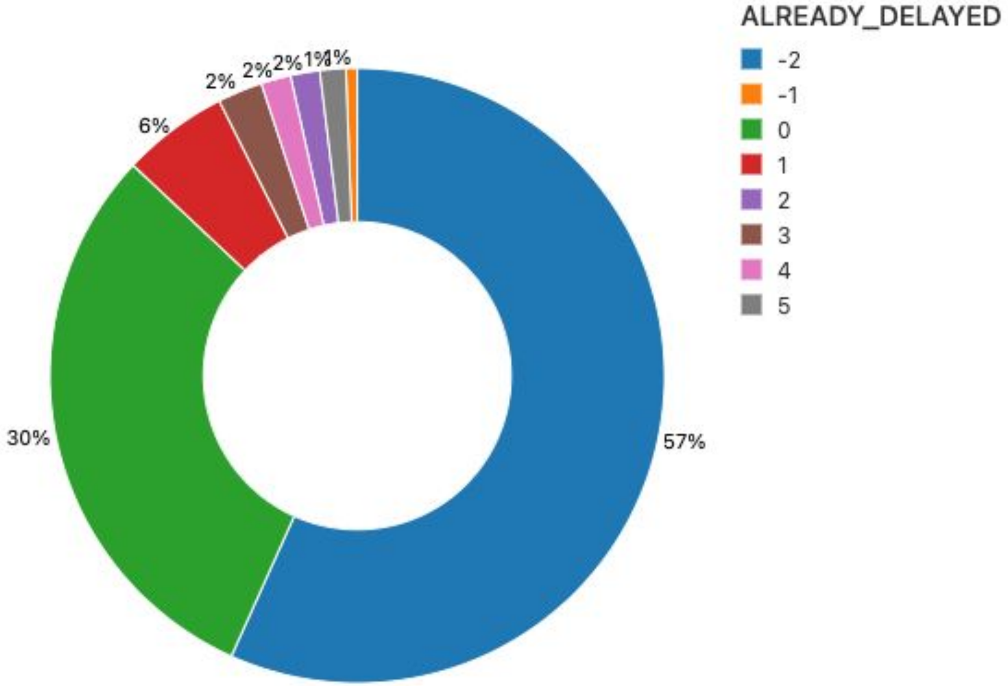
	AIRPORT	PAGERANK_SCORE
1	DEN	0.02833679043953247
2	ORD	0.02720138389532315
3	DFW	0.026113989462048345
4	ATL	0.02094847110411796
5	CLT	0.018122952067459074
6	MSP	0.017373383868555144
7	IAH	0.016012097811575685
8	DTW	0.014814794751604531
9	LAX	0.014294680537843548
10	LAS	0.014139842097311534

Lagged Features

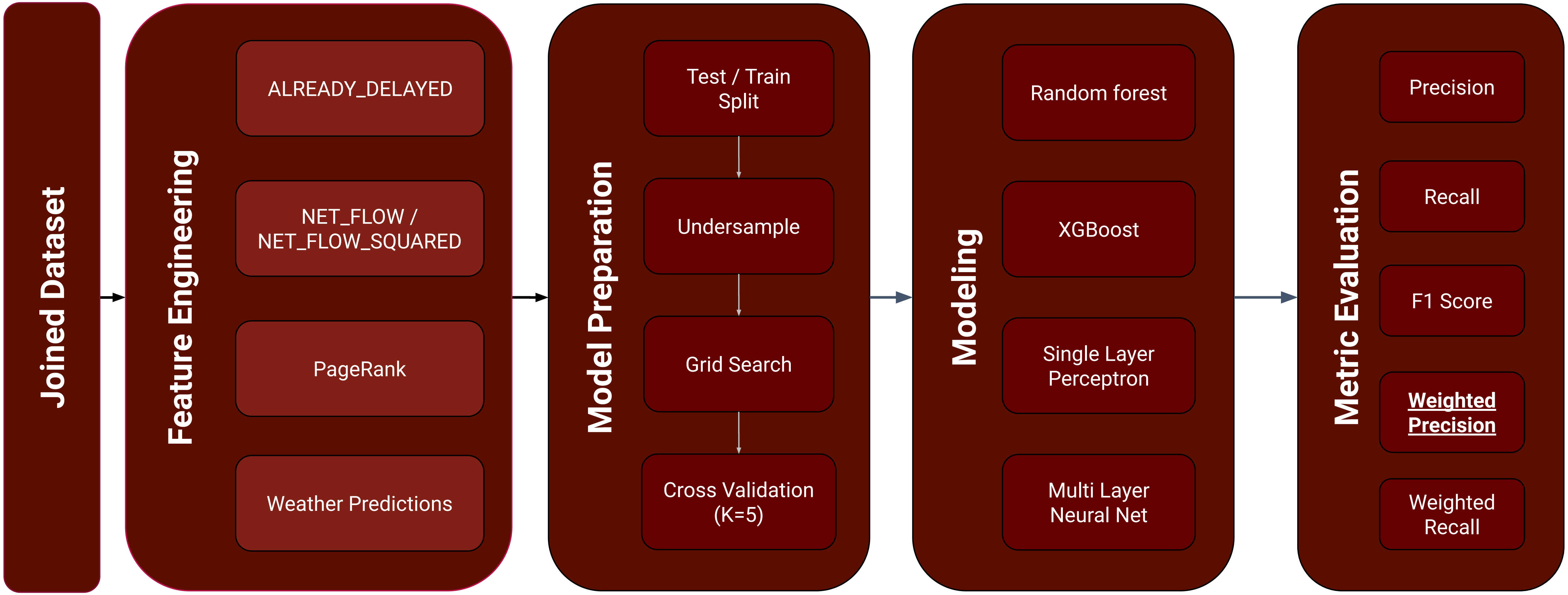
Cascading Delays

- Window function
  - Partition by Tail Number
  - Order by flight date
- Lag Function
  - Quick column addition
- Conditional checks
  - Same day flight
  - No time leakage
- Error handling
  - No prior flight (-2)
  - Attribute Error (-1)

Category	Delay (min)
5	60+
4	30+
3	15+
2	10+
1	1+
0	<= 0



Modeling Pipeline



Results

Algorithm	Time	Weighted Precision	Weighted Recall	Precision (1)	Recall (1)
Random Forest	8.68 minutes	**0.764**	0.806	0.456	0.269
XG Boost	1.56 hours	0.762	0.803	0.446	0.267
Single Hidden Layer Multilayer Perceptron Neural Network	5.85 minutes	0.743	0.773	0.350	0.287
Multiple Hidden Layer Multilayer Perceptron Neural Network	8.24 minutes	0.725	0.612	0.224	0.494

Top 10 Important Features

Random Forest

1. **origin_pred_dew**	64.3
2. origin_pred_wind_speed	5.18
3. origin_pred_wind_direction	4.67
4. origin_pred_wind_direction	4.67
5. origin_pred_wet_bulb_temp	4.5
6. origin_pred_visibility :	4.48
7. origin_pred_pressure :	2.12
8. origin_pred_humidity	2.07
9. origin_pred_sum_precipitation_bulb	1.24
10. origin_pred_dry_bulb	0.7

Conclusion/Future Work

Best Model: Random Forest

Metric	Result
Weighted Precision	0.764
Weighted Recall	0.806
Precision (1)	0.456
Recall (1)	0.269
Time	8.68 minutes

Challenges

- Scalability
- Domain Knowledge
- Collaboration
- Time

Future Work

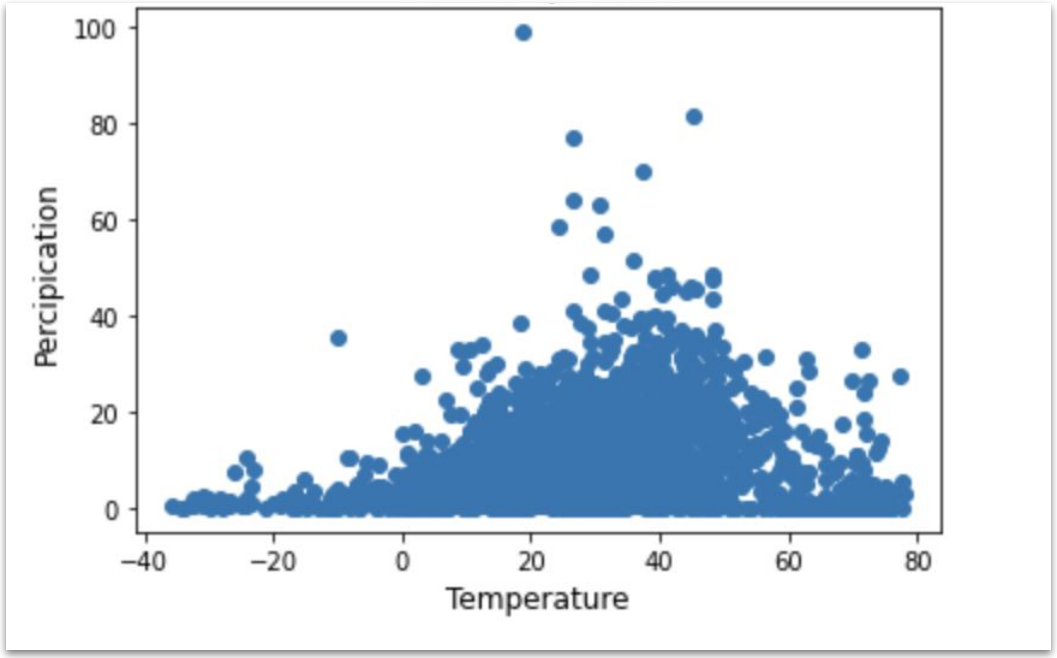
- Upsampling: SMOTE
- Predicting additional features
- Feature Engineering
    - Weather Combinations
    - Weather Movement

Thank You

Appendix

Features

Distribution of Avg. Temperature and Precipitation by Station



Precipitation and temperature are fairly normally distributed across stations

Summary Statistics

- 30M records

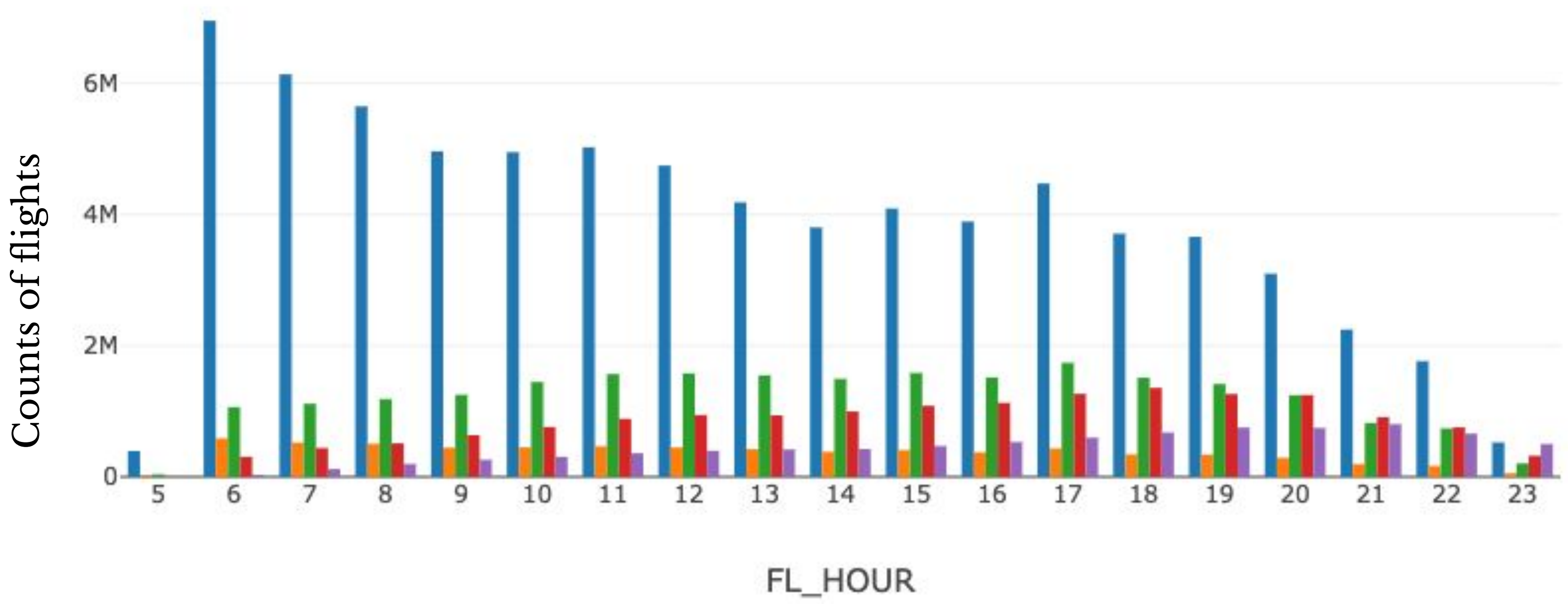
% of Null Values by Feature

Field	% NULL	Field	% NULL		
STATION	0.0%	NormalsHeatingDegreeDay	100.0%	DailyAverageRelativeHumidity	99.7%
DATE	0.0%	ShortDurationEndDate005	100.0%	DailyAverageSeaLevelPressure	99.9%
LATITUDE	0.0%	ShortDurationEndDate010	100.0%	DailyAverageStationPressure	99.9%
LONGITUDE	0.8%	ShortDurationEndDate015	100.0%	DailyAverageWetBulbTemperature	99.7%
ELEVATION	0.8%	ShortDurationEndDate020	100.0%	DailyAverageWindSpeed	99.9%
NAME	0.8%	ShortDurationEndDate030	100.0%	DailyCoolingDegreeDays	99.7%
REPORT_TYPE	0.8%	ShortDurationEndDate045	100.0%	DailyDepartureFromNormalAverageTemperature	99.7%
SOURCE	0.0%	ShortDurationEndDate060	100.0%	DailyHeatingDegreeDays	99.7%
HourlyAltimeterSetting	0.0%	ShortDurationEndDate080	100.0%	DailyMaximumDryBulbTemperature	99.7%
HourlyDewPointTemperature	46.1%	ShortDurationEndDate100	100.0%	DailyMinimumDryBulbTemperature	99.7%
HourlyDryBulbTemperature	17.6%	ShortDurationEndDate120	100.0%	DailyPeakWindDirection	99.7%
HourlyPrecipitation	2.1%	ShortDurationEndDate150	100.0%	DailyPeakWindSpeed	99.7%
HourlyPresentWeatherType	87.1%	ShortDurationEndDate180	100.0%	BackupDistanceUnit	98.3%
HourlyPressureChange	87.1%	ShortDurationPrecipitationValue005	100.0%	BackupElements	98.3%
HourlyPressureTendency	72.4%	ShortDurationPrecipitationValue010	100.0%	BackupElevation	98.3%
HourlyRelativeHumidity	71.4%	ShortDurationPrecipitationValue015	100.0%	BackupEquipment	98.7%
HourlySkyConditions	17.6%	ShortDurationPrecipitationValue020	100.0%	BackupLatitude	98.3%
HourlySeaLevelPressure	47.4%	ShortDurationPrecipitationValue030	100.0%	BackupLongitude	98.7%
HourlyStationPressure	63.8%	ShortDurationPrecipitationValue045	100.0%	BackupName	98.7%
HourlyVisibility	49.2%	ShortDurationPrecipitationValue060	100.0%	WindEquipmentChangeDate	98.2%
HourlyWetBulbTemperature	34.6%	ShortDurationPrecipitationValue080	100.0%	YEAR	93.9%
HourlyWindDirection	50.1%	ShortDurationPrecipitationValue100	100.0%	HTDD	100.0%
HourlyWindGustSpeed	14.3%	ShortDurationPrecipitationValue120	100.0%	NormalsCoolingDegreeDay	100.0%
HourlyWindSpeed	92.8%	ShortDurationPrecipitationValue150	100.0%		
Sunrise	13.2%	ShortDurationPrecipitationValue180	100.0%		
Sunset	99.5%	REM	100.0%		
DailyAverageDewPointTemperature	99.5%	BackupDirection	13.2%		
DailyAverageDryBulbTemperature	99.9%	BackupDistance	98.3%		

A majority of features within the weather dataset are null

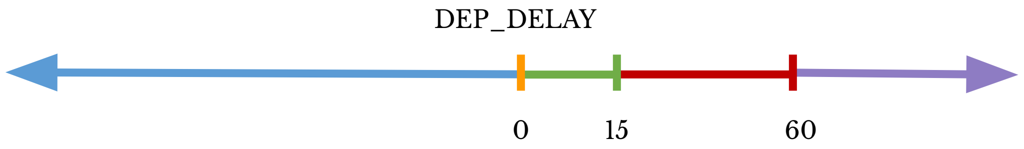
Flights EDA

Categorized Delay by Hour of Day



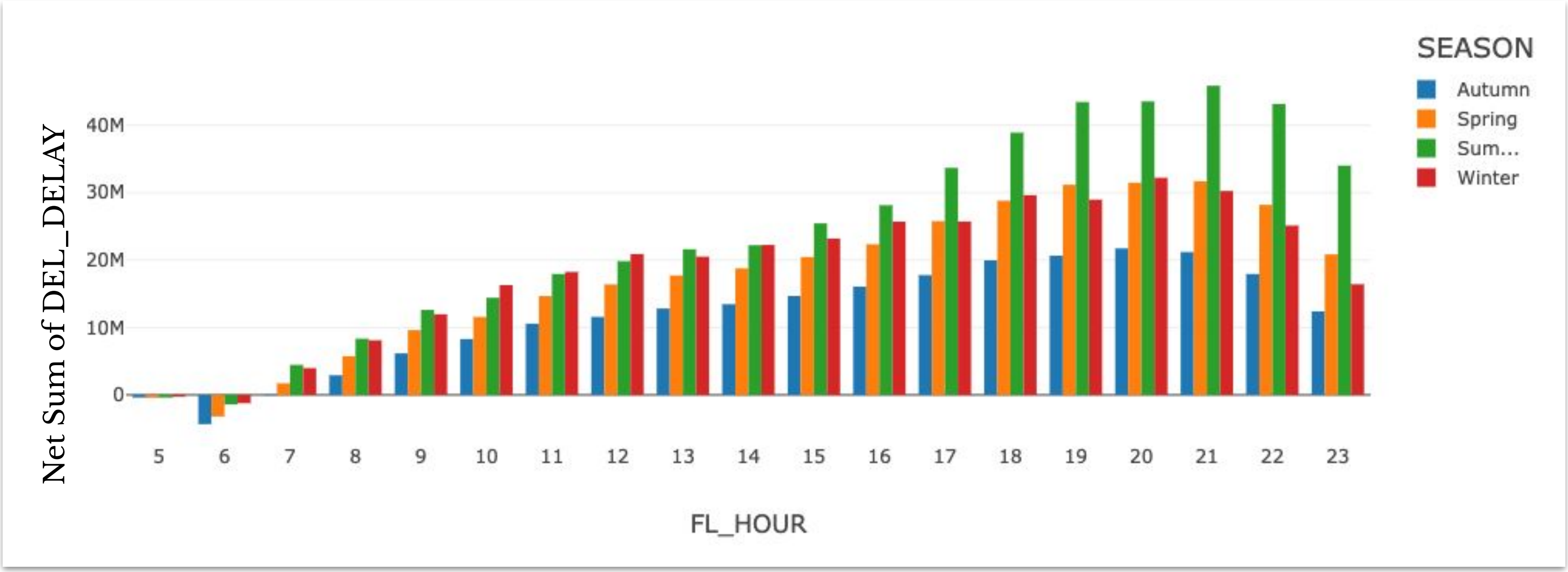
early = DEP_DELAY < 0
ontime = DEP_DELAY = 0
buffer = 0 < DEP_DELAY <= 15
late = 15 < DEP_DELAY <=60
superlate = 60 < DEP_DELAY

Flights are more likely to be on-time earlier in the day, and late later in the day. Time of day is likely to be an important feature while modeling.

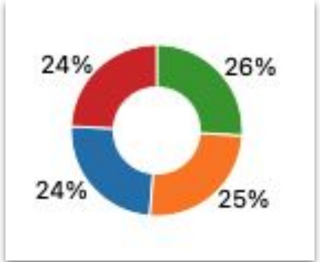


Flights EDA

Delay by Hour of Day and Season

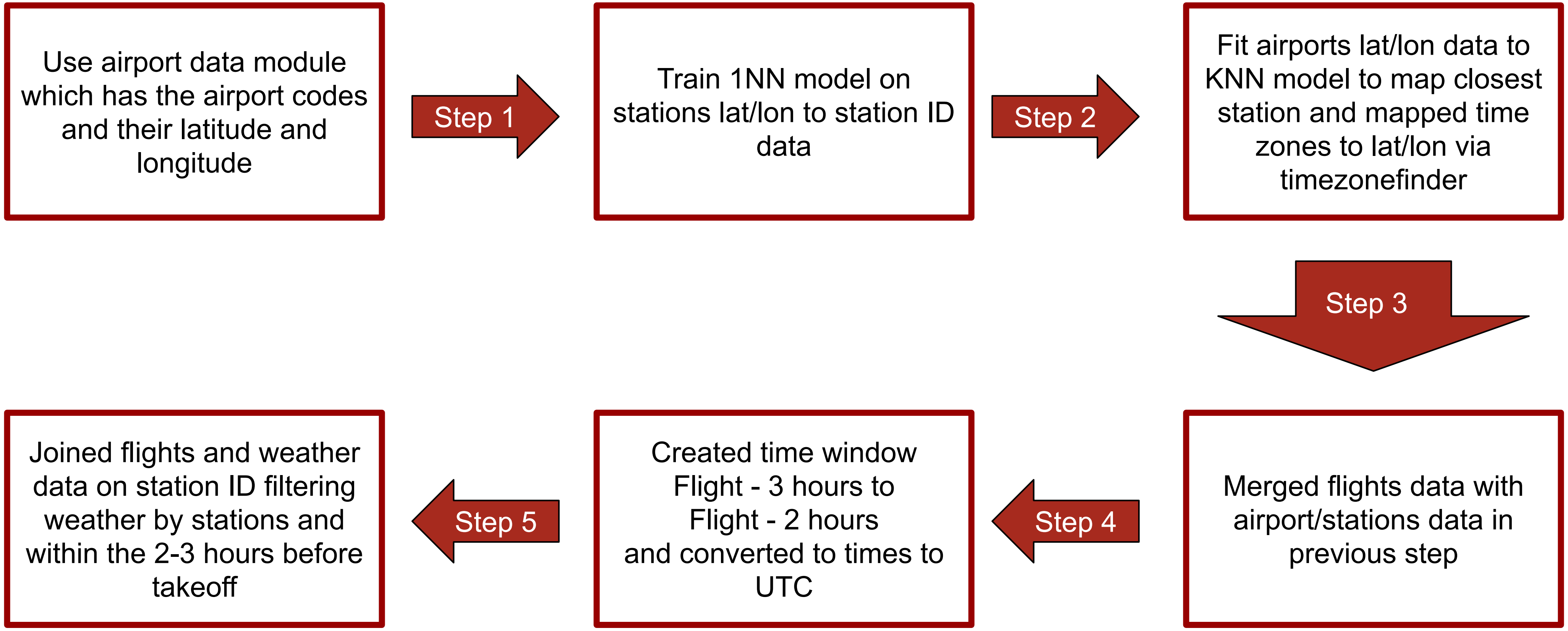


Summer has more delays later in the day. Autumn has the least delays overall. Season is likely to be an important feature while modeling.



Flight distribution throughout the year is fairly even.

Join Process



Data Pipeline

Joins

1. Introduced airport geolocation data in order to join with Stations data
   - Calculated nearest airport using KNN
2. Removed unnecessary features to reduce data size
3. Joined Weather to Airlines on time (minus 2 hours) and origin airport

Feature Engineering

1. Removed unnecessary columns and duplicates
   - ex: dew point dropped when precipitation available
2. Only 2 hours prior used for weather
3. No dimensionality reduction used
4. Features importance

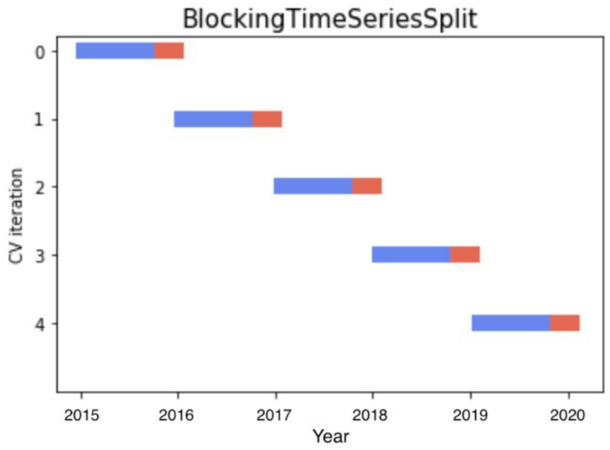
Modeling

1. Cross validation using 1-year cuts
   - Test on random quarter of following year
2. Precision as the main metric
3. Logistic regression as baseline model
4. Look at F1 score(in case data is unbalanced)
5. Decision Trees & Random Forest
6. Parameter optimization(Gridsearch, Input split etc..)

Modeling Improvements

Cross Validation

- Train: Full Year
- Validate: Q1 Following Year



Class Balancing

- Undersampled majority class
- Determined ratio imbalance
- Pyspark SampleBy function
- Not Delayed: ratio% (for each training split)
- Delayed: 100%

Additional Models and Feature Selection

- Logistic Regression Baseline
- Random Forest
- XGBoost
- Single-Layer Neural Net
- Multi-Layer Neural Net

Results

Random Forest

Metric	Result
Weighted Precision	**0.764**
Weighted Recall	0.806
Precision (1)	0.456
Recall (1)	0.269
Time	8.68 minutes

XGBoost

Metric	Result
Weighted Precision	0.71
Weighted Recall	0.67
Precision (1)	0.22
Recall (1)	0.51
Time	1.56 hours

Single Layer Neural Network

Metric	Result
Weighted Precision	0.743
Weighted Recall	0.773
Precision (1)	0.350
Recall (1)	0.287
Time	5.85 minutes

Multilayer Neural Network

Metric	Result
Weighted Precision	0.725
Weighted Recall	0.612
Precision (1)	0.224
Recall (1)	0.494
Time	8.24 minutes

Feature Engineering

Net Flow / Squared Net Flow / PageRank

- Logic
- Implementation

Delay Accumulation

- Window logic:
  - Partition on Tail Number
  - Order by Flight Time
- Ensure flights occurred on same day
- Look at flights that departed 2 hours prior to flight
- 3 hours before - delay maxed at an hour
- 2-3 hours before - delay maxed at minutes between flights

Weather Prediction

- Predicted weather for the flight departure time based on weather inputs 2 hours in advance.
- Used basic Linear Regression with the outcome variables being basic meteorological variables like temperature.

Weather Aggregation

- Calculate averages and sums leading up to the current weather.
- Calculated means for variables like pressure but sums for variables like precipitation.
- Window Functions:
  - Partition by weather station
  - Order by Date descending