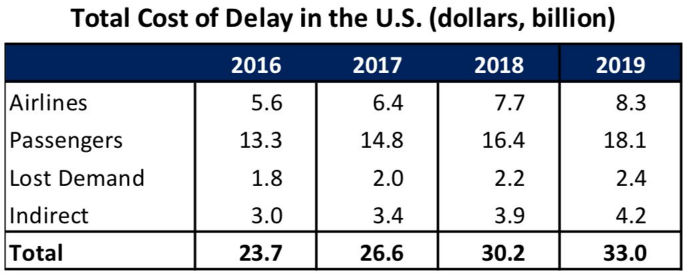
Flight Delays

Final Presentation
Team 25
Jaclyn Andrews, Alyssa Augsburger, John Lee
December 12, 2020

Question Formulation



Total Cost of Delay in the U.S. (dollars, billion)

	2016	2017	2018	2019
Airlines	5.6	6.4	7.7	8.3
Passengers	13.3	14.8	16.4	18.1
Lost Demand	1.8	2.0	2.2	2.4
Indirect	3.0	3.4	3.9	4.2
Total	**23.7**	**26.6**	**30.2**	**33.0**

Question:

Given flight and weather information known two hours ahead of planned departure time, will a flight depart on time (within 15 minutes of scheduled departure) or will it be delayed or cancelled?

Evaluation:

F1 Score - to minimize both false positive and negatives

State of the Art (on a subset of data):

F1 Score of 0.85

https://www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf

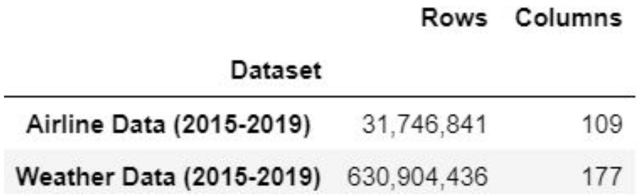
Data Introduction

Flights Data:

- Reporting Carrier On-Time Performance (Bureau of Transportation Statistics)
- 2015 through 2019 U.S. Flights Data

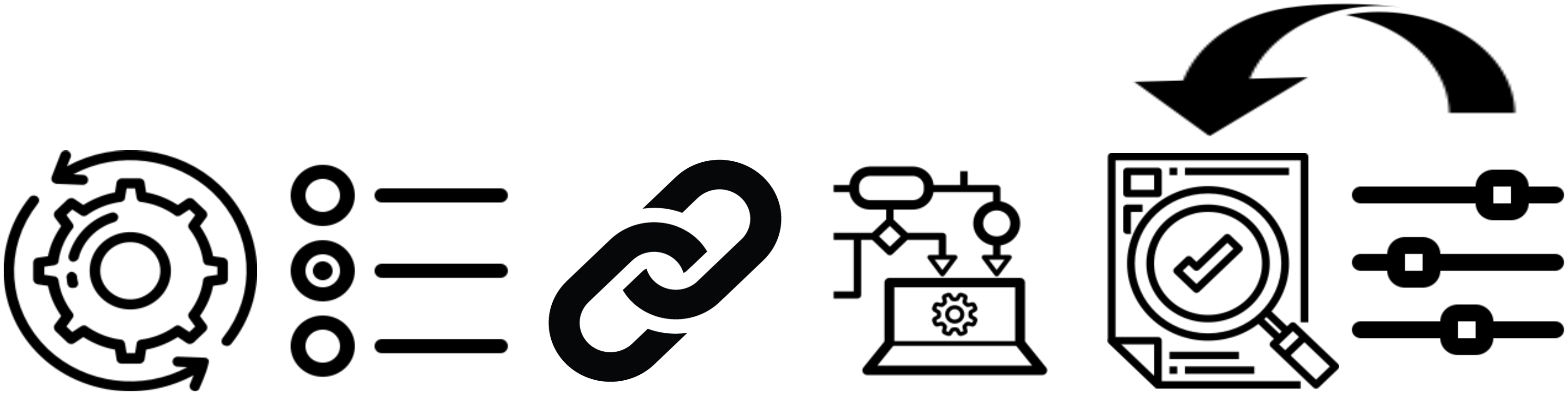
Weather Data:

- National Oceanic and Atmospheric Administration Repository
- 2015 through 2019 Weather Data



	Rows	Columns
Dataset		
Airline Data (2015-2019)	31,746,841	109
Weather Data (2015-2019)	630,904,436	177

Our Process



EDA + Data Pre-Processing

Feature Engineering

Join Data

Train Machine Learning Models

Evaluate Results

Fine Tune Models

EDA: Large Dataframes

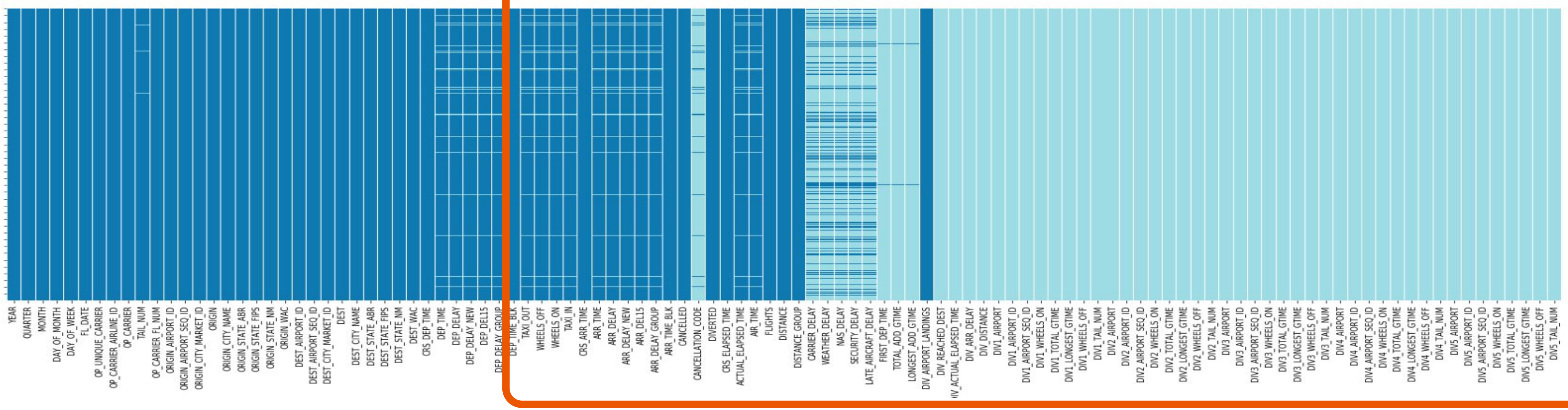
Challenge: High computational complexity
Solution: Filter and sub-setting (asw)

Challenge: Handling nulls
Solution: Drop training nulls & impute test nulls

Viable features

Taxi, arrival, in flight details

Delay and diversion details, mostly null

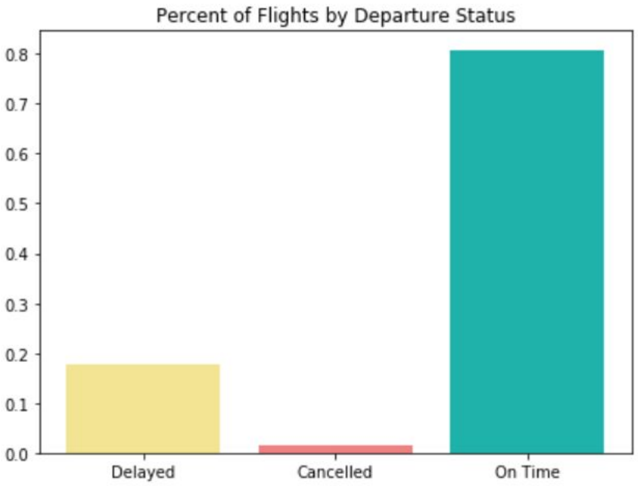


Code:

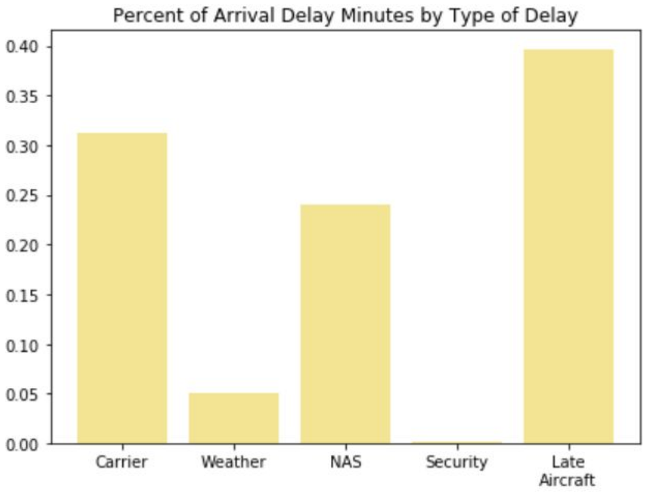
```
sns.heatmap(airlines_3m.toPandas().isnull(), cmap='tab20', cbar=False);
```

EDA: Outcome Variable

Challenge: Unbalanced data
Solution: Sampling weights

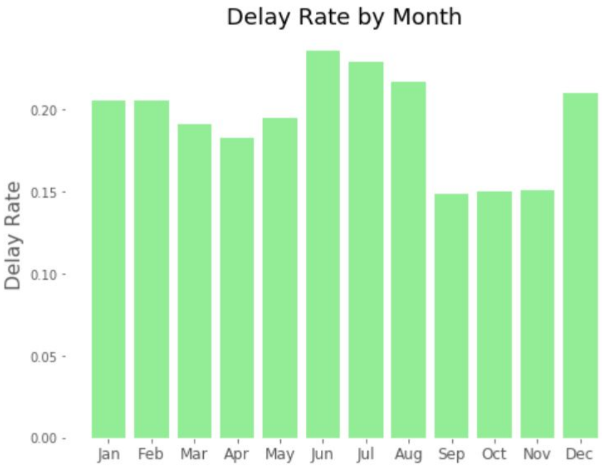


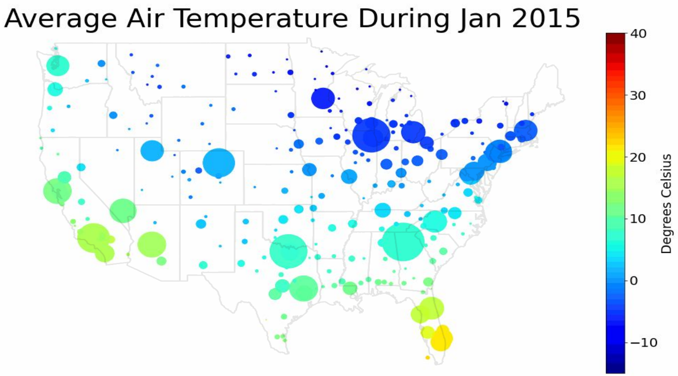
Challenge: Various separate causes of delay
Solution: Customized feature creation

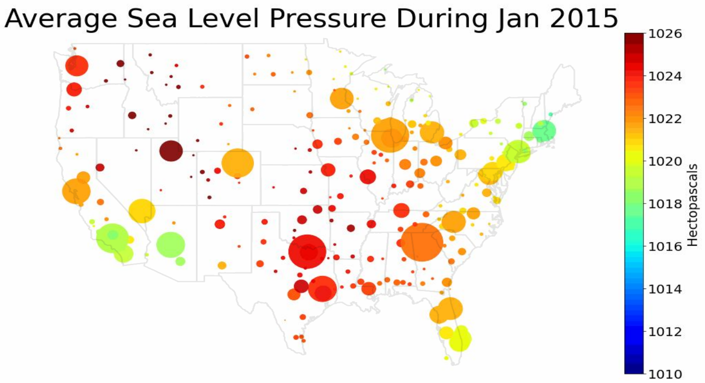


EDA: Seasonality

Challenge: Capturing the variances in seasonality
Solution: Split data into train/validation/test by year

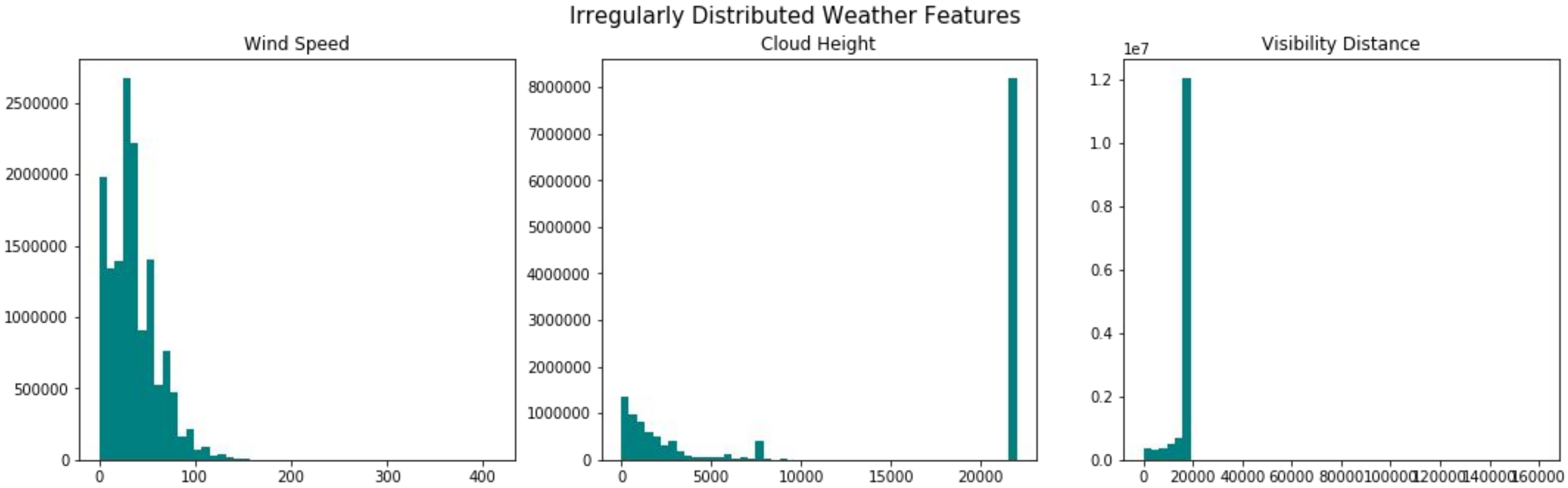




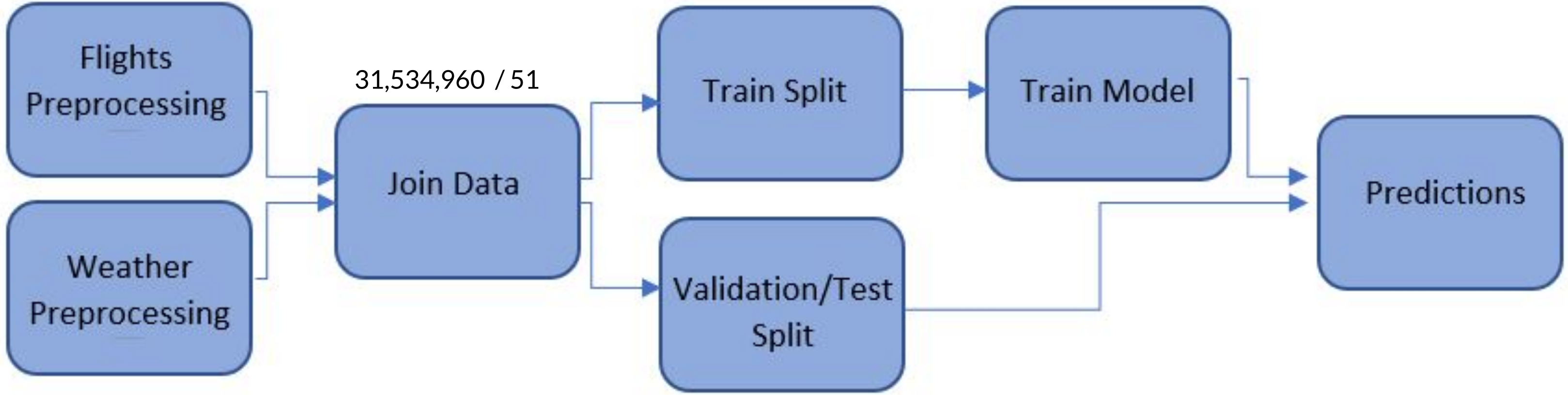


EDA: Weather Data Abnormalities

Challenge: Data is not normally distributed
Solution: Binned data on splits



Data Processing: Pipeline



31,746,841 / 109

31,737,432 / 22

31,534,960 / 51

14,771,645 / 36

630,904,436 / 177

22,283,287 / 21

7,170,842 / 36
7,307,855 / 36

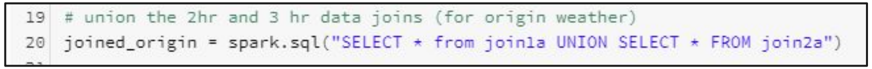
Legend: Rows / Columns

Data Processing: Data Join

Flights & Weather Data Join

- Flights airport code (IATA <-> ICAO)
- Flights datetime transformation
- Flights timezone adjustment
- Join on data 2-hours prior

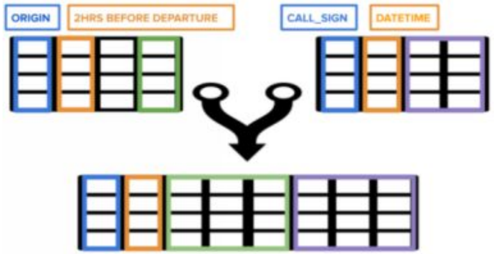
```
# join 2 hr data (for origin weather)
join1a = spark.sql("SELECT * FROM a_tt INNER JOIN w_tt ON (a_tt.CRS_DEP_TIME_2HR_HR = w_tt.DATE_HR AND a_tt.origin_icao_code = w_tt.AIRPORT)")
dbutils.fs.rm("dbfs:/mnt/mids-w261/team_25/join_data_folder/join1a", True)
join1a.write.parquet("dbfs:/mnt/mids-w261/team_25/join_data_folder/join1a")
join1a = spark.read.option("header", "true").parquet("dbfs:/mnt/mids-w261/team_25/join_data_folder/join1a/part-00*.parquet")
join1a.registerTempTable("join1a")
```



```
# union the 2hr and 3 hr data joins (for origin weather)
joined_origin = spark.sql("SELECT * from join1a UNION SELECT * FROM join2a")
```

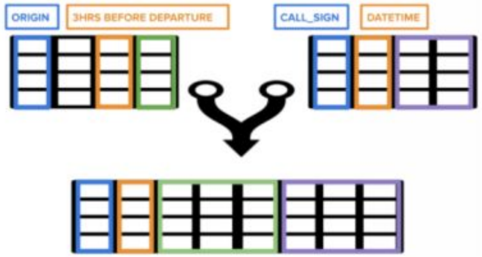
Step 1: Inner Join

Join on airport and time bucket 2 hours prior to departure.



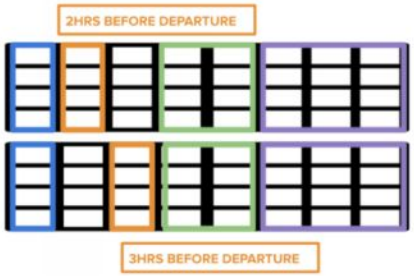
Step 2: Inner Join

Join on airport and time bucket 3 hours prior to departure.



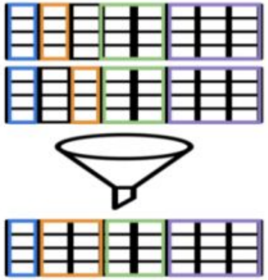
Step 3: Union

Stack two joins on top of each other to make one set of flights with all weather readings from 2 and 3 hours prior to departure.



Step 4: Filter & Group By

Filter out weather readings later than 2 hours prior to departure. Group by flight and select latest remaining weather reading.



Feature Engineering: Selection and Transformation

		Variable Type	
		Categorical One Hot Encoded for Regression	Numerical Normalized for Regression
Data Source	Airlines	Month Day of week Carrier Origin airport Origin state Destination airport Destination state Departure time (4 hr bins) First departure (Yes/No binary)	Elapsed Time Distance
	Weather*	Wind speed (Beaufort scale bins) Ceiling height (Limited/Unlimited binary) Visibility distance (+/- 10 miles binary) Visibility variability	Latitude Longitude Elevation Temperature Dew point Sea level pressure

**All weather variables included for both origin and destination weather stations 2 hours prior to departure*

● Altered from original numerical format to bins

Feature Engineering: Creation

Delay Propagation

(Tail # Previous Delay)

Derived from: flight delay data available 2 hours prior to flight departure.
Values: 0 (no prev. delay), 1 (prev. delay)
Missing: Default value of no previous delay

Flight Schedule

(# of flights per day)

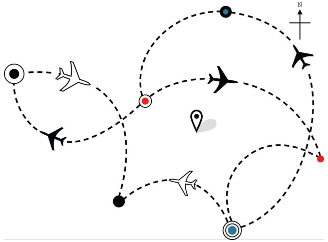
Derived from: CRS flight schedule*
Values: numeric, # of flights
Missing:

**assumption: schedule is provided and determined at least a full day in advance.*

Airport Importance

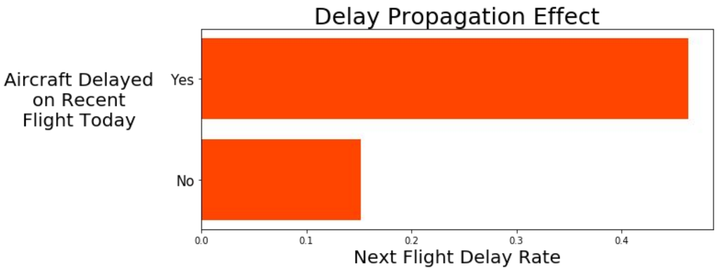
(PageRank)

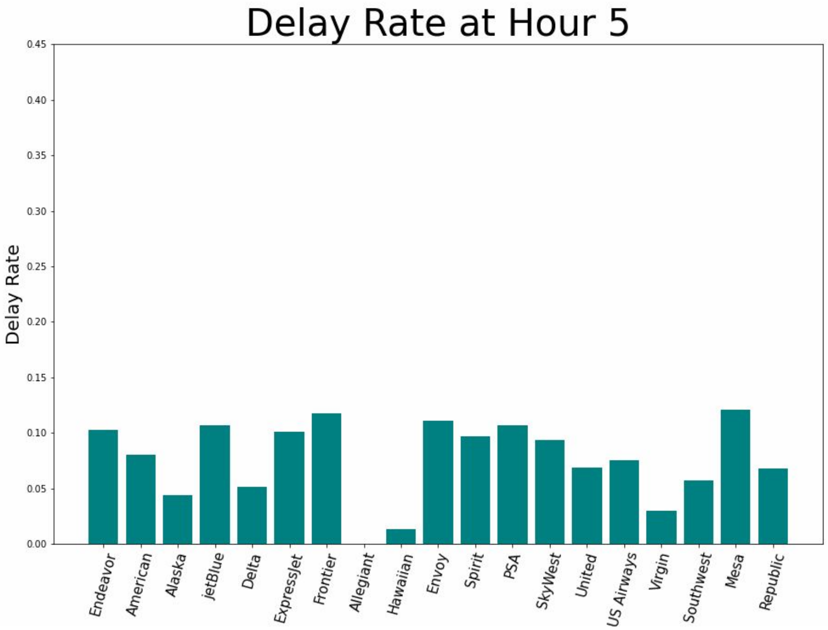
Derived from: Directed flight path graph weighted by number of flights along the path.
Values: rank between 0 and 1
Missing: imputed with 1/(# of airports in training data)



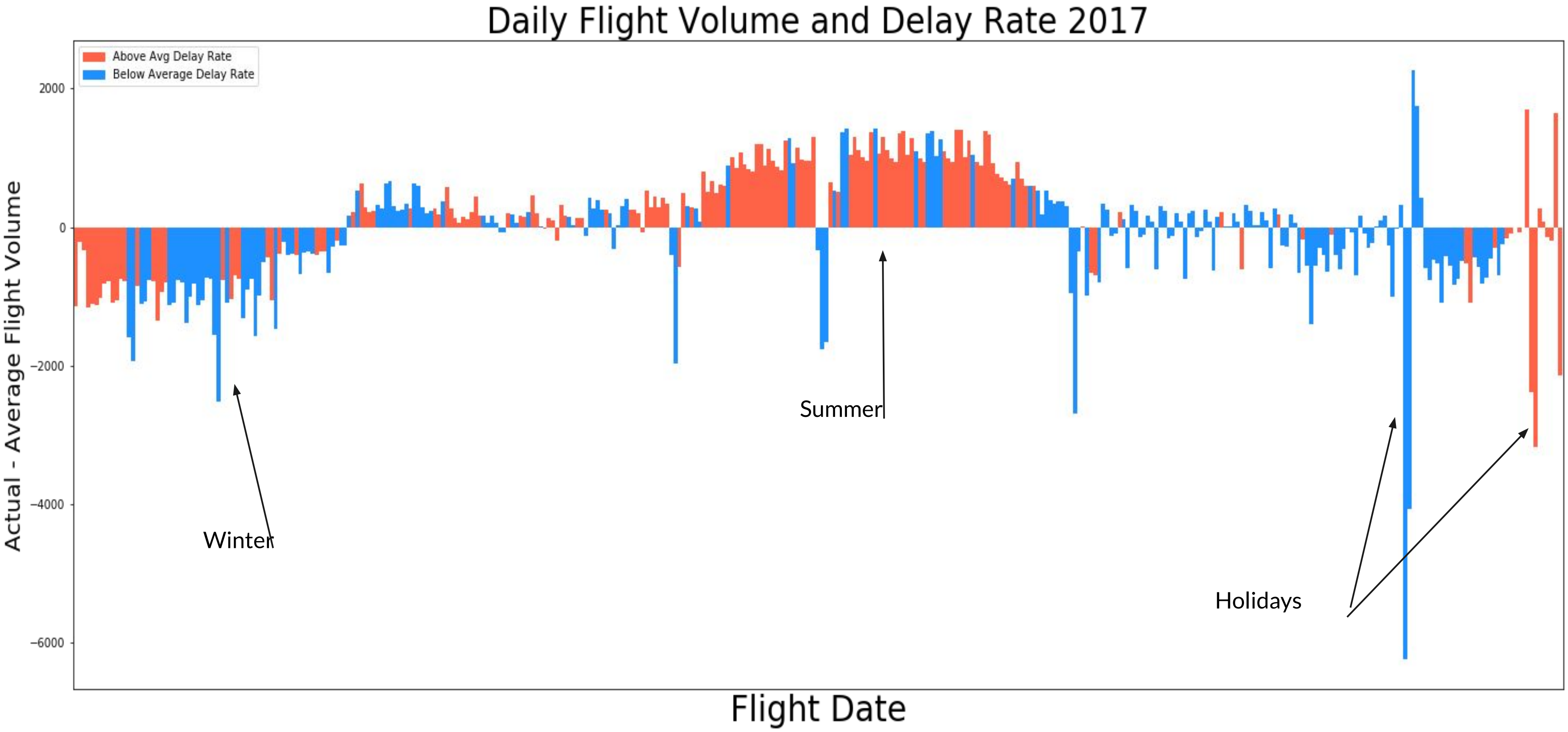
Feature Engineering: Delay Propagation

- Delay rate increases throughout the day across all airlines
- Major cause of delay is late aircrafts
- Features to include:
    - Time of day
    - Binary first flight of day
    - Aircraft’s last flight delay status

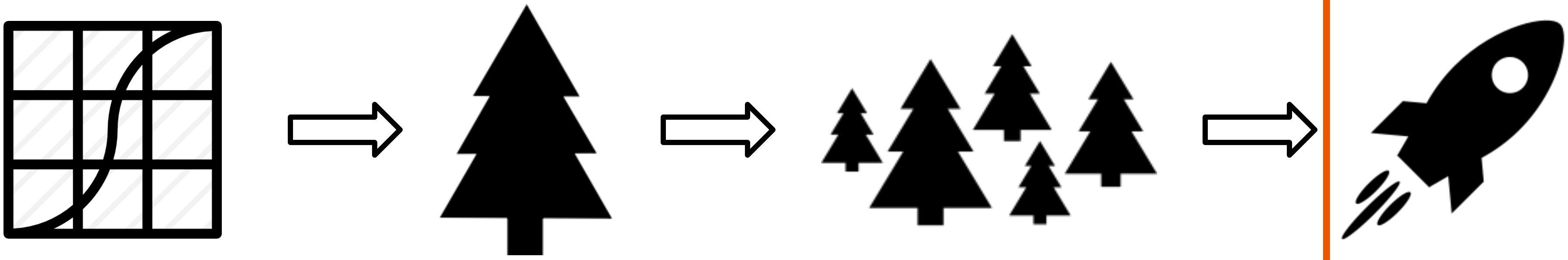




Feature Engineering: Flight Schedule



Algorithm Exploration: Baseline



Logistic Regression

F1-Score: 0.372

Decision Tree

F1-Score: 0.374

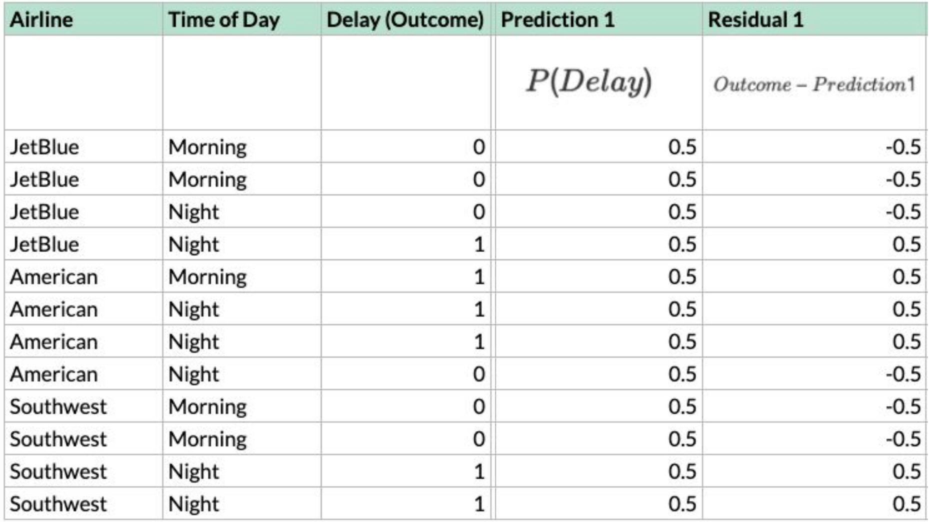
Random Forest

F1-Score: 0.402

Gradient Boosted Tree

F1-Score: 0.413

Algorithm Implementation: Toy Example

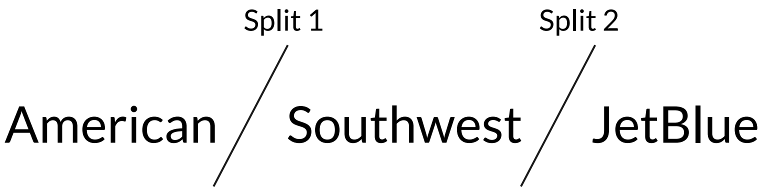


Airline	Time of Day	Delay (Outcome)	Prediction 1	Residual 1
			$P(Delay)$	$Outcome - Prediction1$
JetBlue	Morning	0	0.5	-0.5
JetBlue	Morning	0	0.5	-0.5
JetBlue	Night	0	0.5	-0.5
JetBlue	Night	1	0.5	0.5
American	Morning	1	0.5	0.5
American	Night	1	0.5	0.5
American	Night	1	0.5	0.5
American	Night	0	0.5	-0.5
Southwest	Morning	0	0.5	-0.5
Southwest	Morning	0	0.5	-0.5
Southwest	Night	1	0.5	0.5
Southwest	Night	1	0.5	0.5

$$P(Delay)_{JetBlue} = 0.75$$

$$P(Delay)_{American} = 0.25$$

$$P(Delay)_{Southwest} = 0.5$$



Split 1

Split 2

American

Southwest

JetBlue

Algorithm Implementation: Toy Example

$$\text{Gini index}_{Time=\underline{Morning}} = 1 - \sum_{i=1}^{n} {p_i}^2 = 1 - P(Delay)^2 - P(No\,delay)^2 = 1 - \frac{1}{5}^2 - \frac{4}{5}^2 = \underline{0.320}$$

$$\text{Gini index}_{Time=Night} = 1 - \frac{5}{7}^2 - \frac{2}{7}^2 = 0.408$$

$$\text{Weighted gini index}_{Time} = \frac{5}{12} \times \underline{0.320} + \frac{7}{12} \times 0.408 = \boxed{0.371}$$

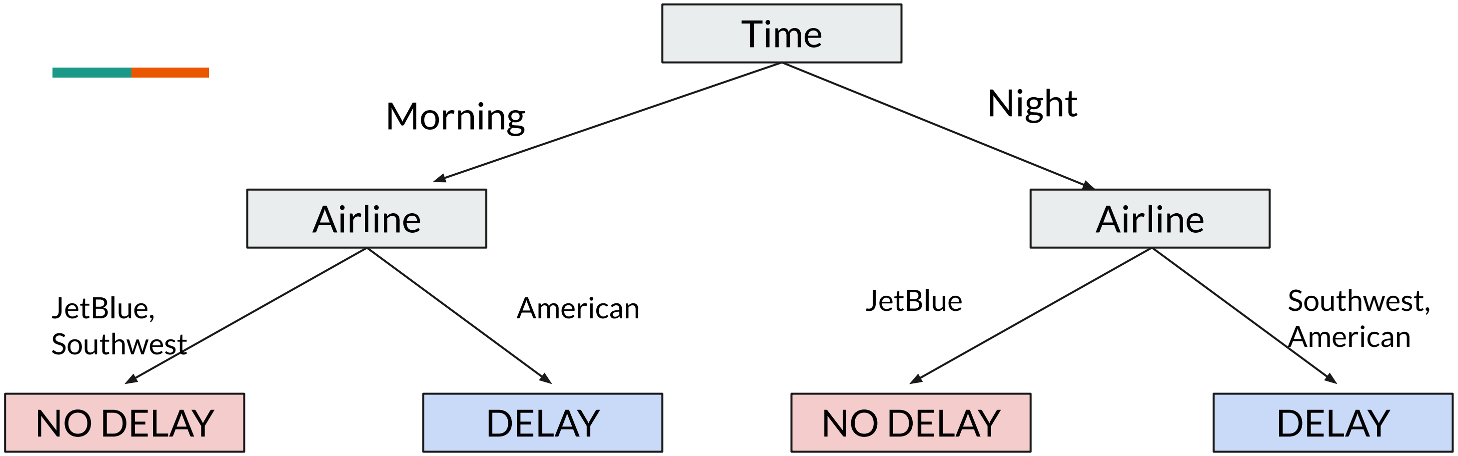
Airline	Time of Day	Delay (Outcome)
JetBlue	Morning	0
JetBlue	Morning	0
JetBlue	Night	0
JetBlue	Night	1
American	Morning	1
American	Night	1
American	Night	1
American	Night	0
Southwest	Morning	0
Southwest	Morning	0
Southwest	Night	1
Southwest	Night	1

ROOT	
Split Point	**Gini index**
Time (weighted)	0.371
Airline Split1 (weighted)	0.438
Airline Split2 (weighted)	0.438

MORNING	
Split Point	**Gini index**
Airline Split1 (weighted)	0.267
Airline Split2 (weighted)	0.000

NIGHT	
Split Point	**Gini index**
Airline Split1 (weighted)	0.371
Airline Split2 (weighted)	0.405

Algorithm Implementation: Toy Example



Time

Morning

Night

Airline

Airline

JetBlue, Southwest

American

JetBlue

Southwest, American

NO DELAY

DELAY

NO DELAY

DELAY

-0.5, -0.5, -0.5, -0.5

0.5

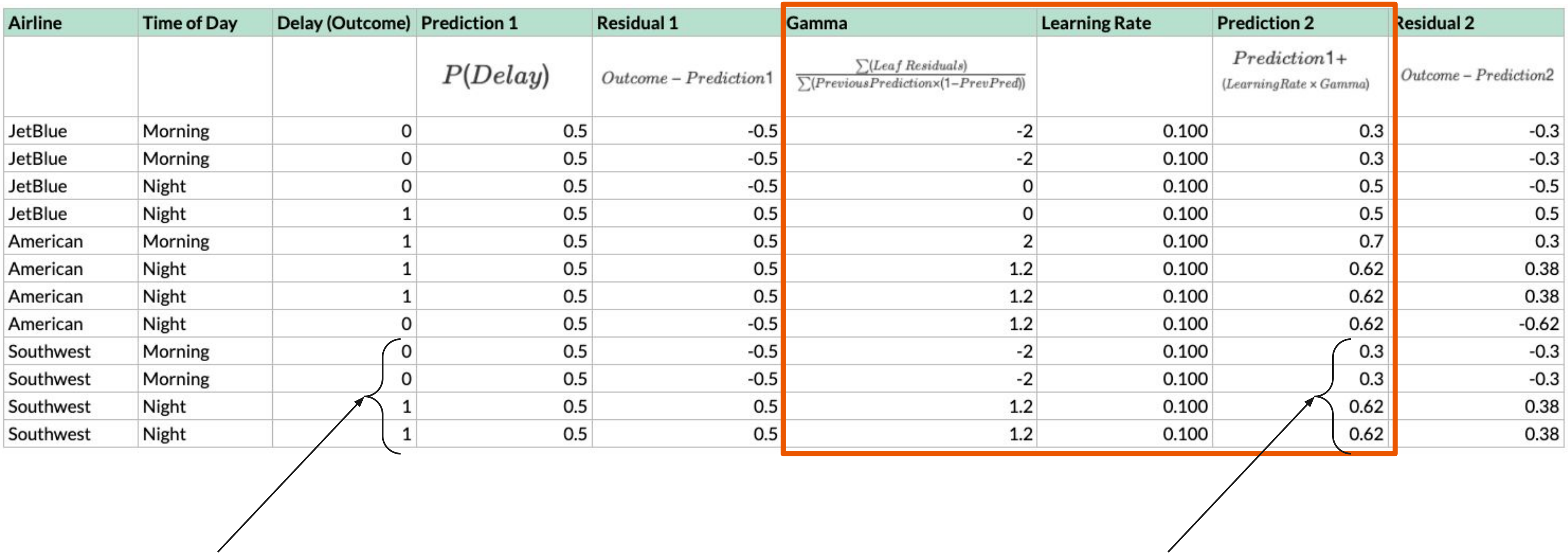
-0.5, 0.5

-0.5, 0.5, 0.5, 0.5, 0.5

$$\text{Gamma} = \frac{\sum(Leaf\ Residuals)}{\sum(PreviousPrediction \times (1 - PrevPred))} = \frac{0.5}{0.5(1-0.5)} = 2$$

-2	2	0	1.2

Algorithm Implementation: Toy Example



Airline	Time of Day	Delay (Outcome)	Prediction 1	Residual 1	Gamma	Learning Rate	Prediction 2	esidual 2
			$P(Delay)$	$Outcome - Prediction1$	$\frac{\sum(Leaf\ Residuals)}{\sum(PreviousPrediction \times (1 - PrevPred))}$		$Prediction1 + (LearningRate \times Gamma)$	$Outcome - Prediction2$
JetBlue	Morning	0	0.5	-0.5	-2	0.100	0.3	-0.3
JetBlue	Morning	0	0.5	-0.5	-2	0.100	0.3	-0.3
JetBlue	Night	0	0.5	-0.5	0	0.100	0.5	-0.5
JetBlue	Night	1	0.5	0.5	0	0.100	0.5	0.5
American	Morning	1	0.5	0.5	2	0.100	0.7	0.3
American	Night	1	0.5	0.5	1.2	0.100	0.62	0.38
American	Night	1	0.5	0.5	1.2	0.100	0.62	0.38
American	Night	0	0.5	-0.5	1.2	0.100	0.62	-0.62
Southwest	Morning	0	0.5	-0.5	-2	0.100	0.3	-0.3
Southwest	Morning	0	0.5	-0.5	-2	0.100	0.3	-0.3
Southwest	Night	1	0.5	0.5	1.2	0.100	0.62	0.38
Southwest	Night	1	0.5	0.5	1.2	0.100	0.62	0.38

Tuning

Individual Feature Inclusion:

- ✓ PageRank
- ✓ Delay propagation
- ✓ Destination weather
- ✗ Flight schedule

Hyperparameters:

- ✗ Maximum depth
- ✗ Maximum iterations
- ✓ Minimum instances per node

Other:

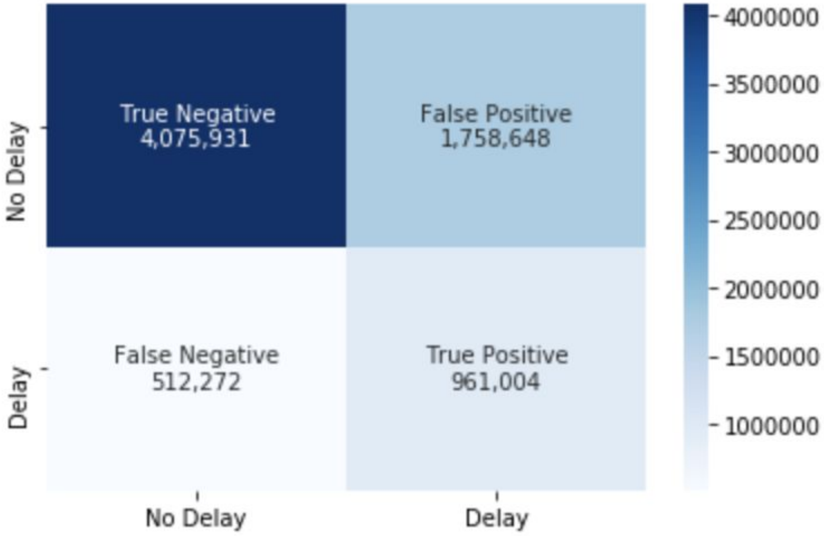
- ✓ One hot encoding
- ✓ Top 10 features from other models
- ✓ Principal component analysis

✓ Included in final model

✗ Not included in final model

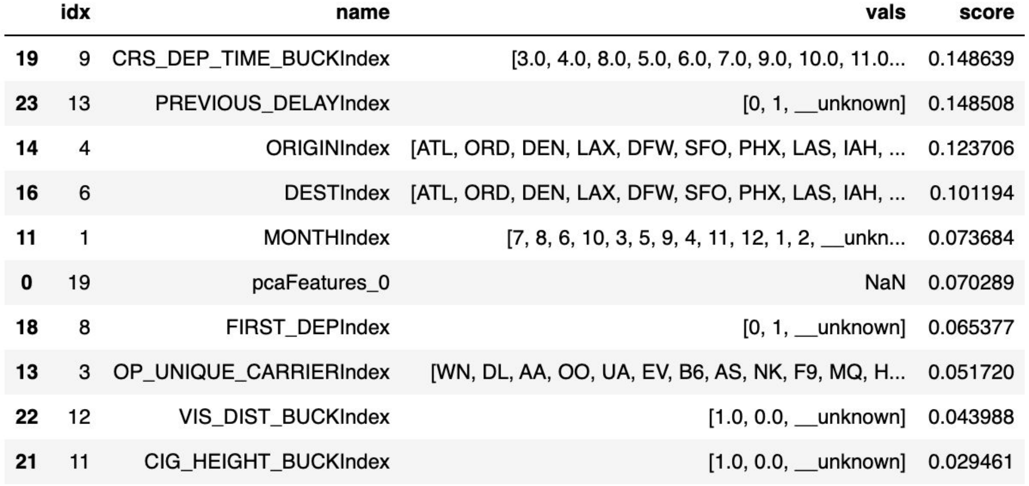
Final Algorithm

- Gradient Boosted Tree
    - PCA with 10 components
    - Maximum depth = 6
    - Maximum iterations = 20
    - Minimum instances per node = 1
- Results
    - **F1 Score: 0.458**
    - **Recall: 0.652**
    - Precision: 0.353
    - Accuracy: 0.689
    - AUC: 0.743



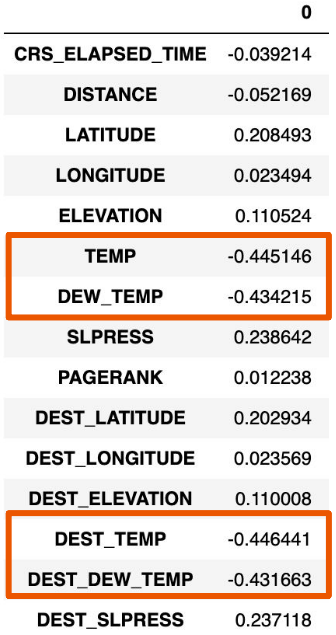
Feature Importance

Top 10 Important Features



	idx	name	vals	score
19	9	CRS_DEP_TIME_BUCKIndex	[3.0, 4.0, 8.0, 5.0, 6.0, 7.0, 9.0, 10.0, 11.0...	0.148639
23	13	PREVIOUS_DELAYIndex	[0, 1, __unknown]	0.148508
14	4	ORIGINIndex	[ATL, ORD, DEN, LAX, DFW, SFO, PHX, LAS, IAH, ...	0.123706
16	6	DESTIndex	[ATL, ORD, DEN, LAX, DFW, SFO, PHX, LAS, IAH, ...	0.101194
11	1	MONTHIndex	[7, 8, 6, 10, 3, 5, 9, 4, 11, 12, 1, 2, __unkn...	0.073684
0	19	pcaFeatures_0	NaN	0.070289
18	8	FIRST_DEPIndex	[0, 1, __unknown]	0.065377
13	3	OP_UNIQUE_CARRIERIndex	[WN, DL, AA, OO, UA, EV, B6, AS, NK, F9, MQ, H...	0.051720
22	12	VIS_DIST_BUCKIndex	[1.0, 0.0, __unknown]	0.043988
21	11	CIG_HEIGHT_BUCKIndex	[1.0, 0.0, __unknown]	0.029461

pcaFeatures_0



	0
CRS_ELAPSED_TIME	-0.039214
DISTANCE	-0.052169
LATITUDE	0.208493
LONGITUDE	0.023494
ELEVATION	0.110524
TEMP	-0.445146
DEW_TEMP	-0.434215
SLPRESS	0.238642
PAGERANK	0.012238
DEST_LATITUDE	0.202934
DEST_LONGITUDE	0.023569
DEST_ELEVATION	0.110008
DEST_TEMP	-0.446441
DEST_DEW_TEMP	-0.431663
DEST_SLPRESS	0.237118

Conclusion

- Performance and scalability
    - Gradient boosted trees can’t fully be parallelized
- Future work
    - Additional engineered features
    - Hyperparameter tuning
    - Split data at very beginning of pipeline

Bibliography

- Lukacs, Michael. “Cost of Delay Estimates.” FAA, 2019, www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf.
- “OST_R: BTS: Transtats.” BTS, 2020, www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120.
- “Land-Based Datasets and Products.” National Climatic Data Center, 2020, www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets.
- Belcastro, Loris, et al. Using Scalable Data Mining for Predicting Flight Delays. Dec. 2014.
- Chakrabarty, Navoneel. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.