



On Time or Delayed?

Final Project, Team 30

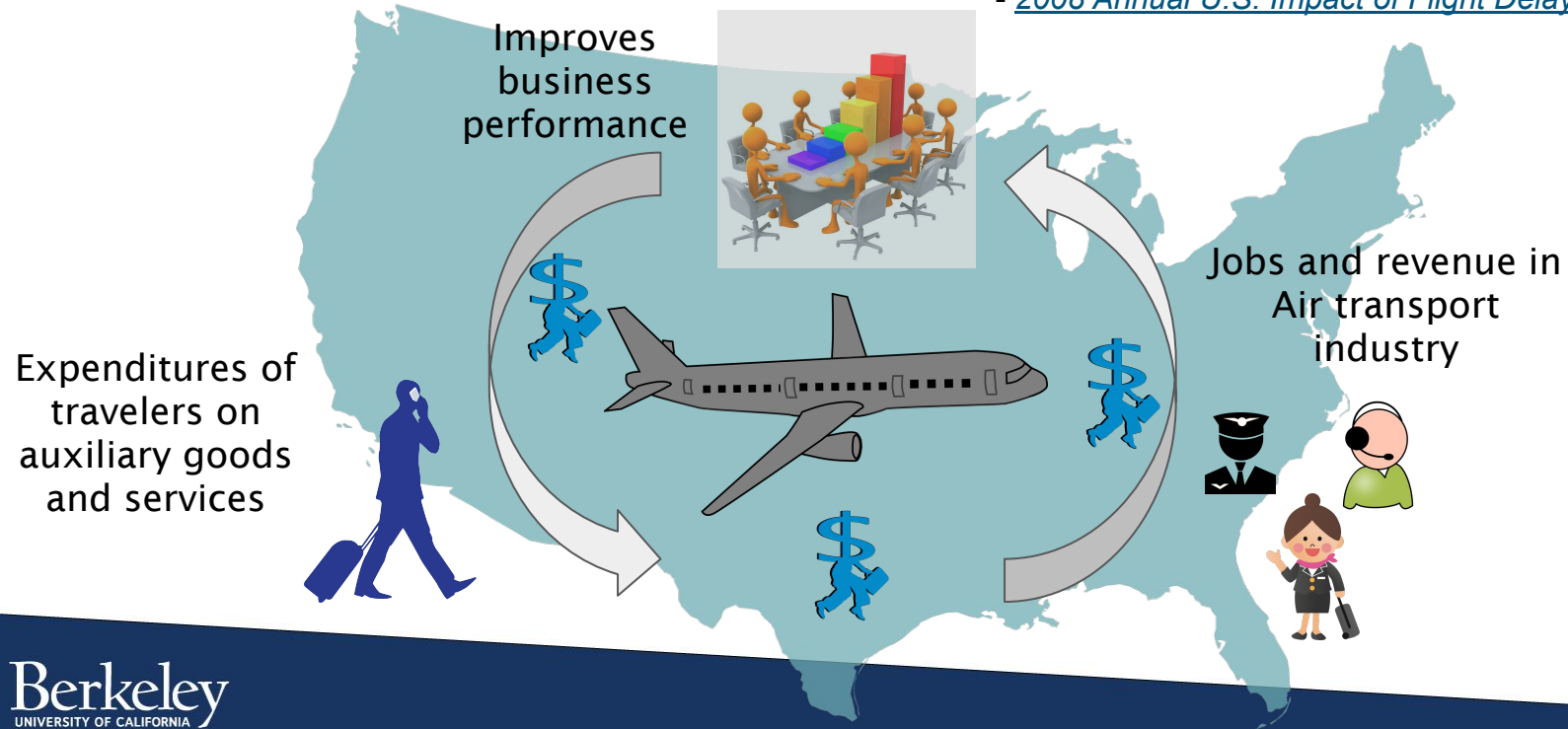
Chitra Agastya, Patti Degner, Gunnar Mein

W261, Sec 5

Motivation

“growing delays threaten the competitiveness of the U.S. in the world economy by limiting the ability of the air transport system to serve the needs of the U.S. economy”

- [2008 Annual U.S. Impact of Flight Delays report](#)



Question and Outcome



Can a **4-hour window of weather data** at origin and destination, along with key **flight performance indicators** help machines foresee a *'predictable'* delay of 15 min or more, two hours before the scheduled takeoff?

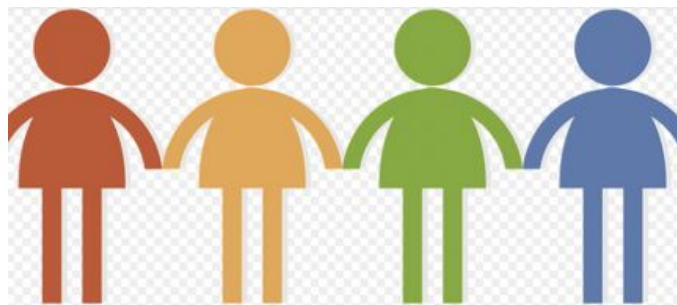






A photograph of a flight information display board. A green rectangular box highlights the word "ON TIME" in green capital letters. The board lists flight details including flight numbers, destinations, times, and statuses.

Flight Number	Destination	Time	Status
2106	D01	11:05am	On Time
4547	B12	11:15am	Boarding
780	C03	1:30pm	On Time
4649	E13	11:05am	Boarding
5296	E13	3:00pm	On Time
6729	D19	2:00pm	On Time
7383	E70	11:00 am	On Time
7466	B7	11:10am	On Time
		11:09am	On Time

Evaluation Metrics: Perspectives

Passengers



Prediction	On Time		Delayed
True label			
On Time	No reaction		"What? I missed my flight?" 
Delayed	"This app is useless"		"I saved a few minutes" 

Prediction	On Time		Delayed
True label			
On Time			
Delayed			  BONUS

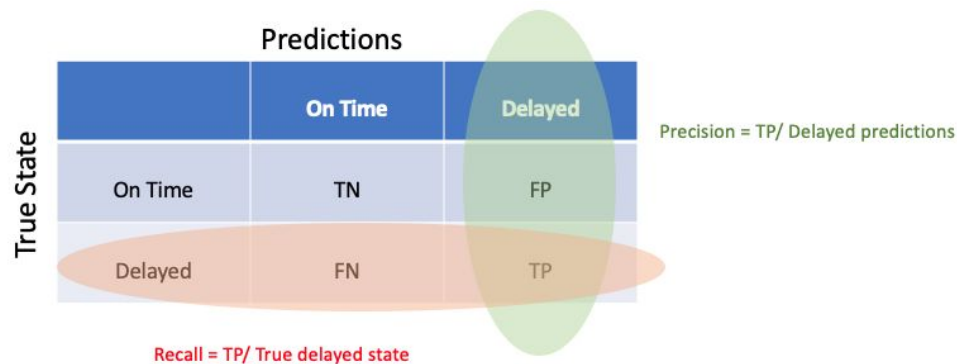


Evaluation Metrics (cont.)

- Plausible value for airlines, other businesses that can react *statistically*
- Our objective is to **maximize Precision**

$$\text{Precision} = TP / (TP + FP)$$

- Look at raw numbers too!



Data Overview



Airport Data

Flight information (time of travel, airports, carrier, delay, other performance indicators)

[OST_R](#) | [BTS](#) | [Transtats](#)



Weather Data

Wind angle/speed, ceiling, visibility, temperature, dewpoint, pressure

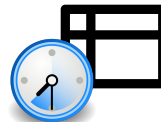
[National Oceanic and Atmospheric Administration repository](#)



Weather Station Data

Name, coordinates of a weather station

<http://dss.ucar.edu/datasets/ds353.4/inventories/station-list.html>

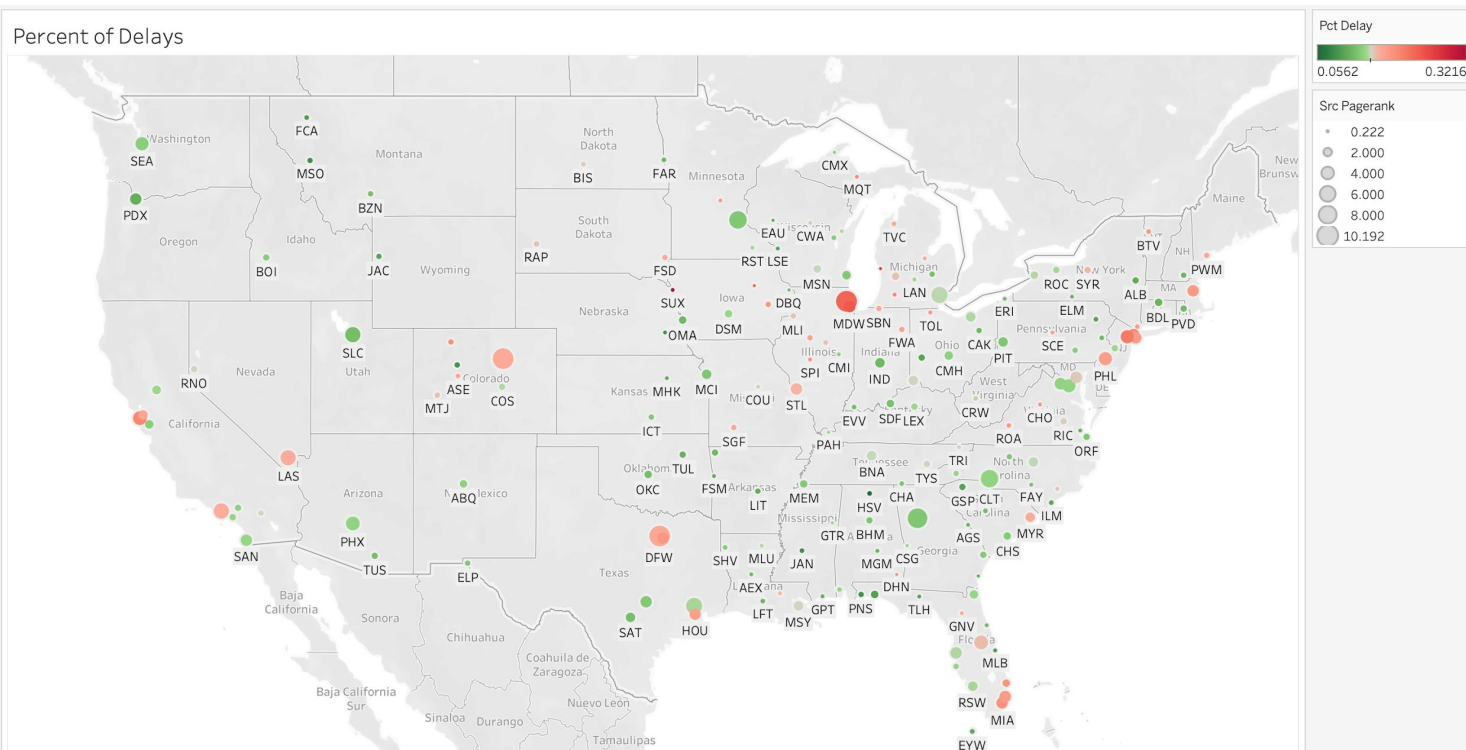


Timezone Data

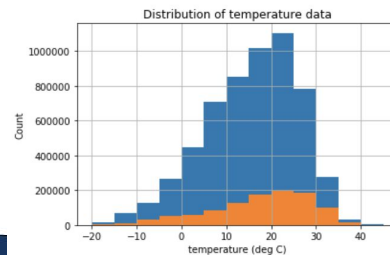
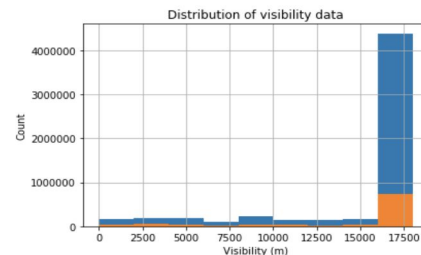
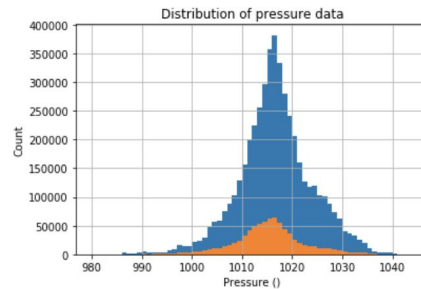
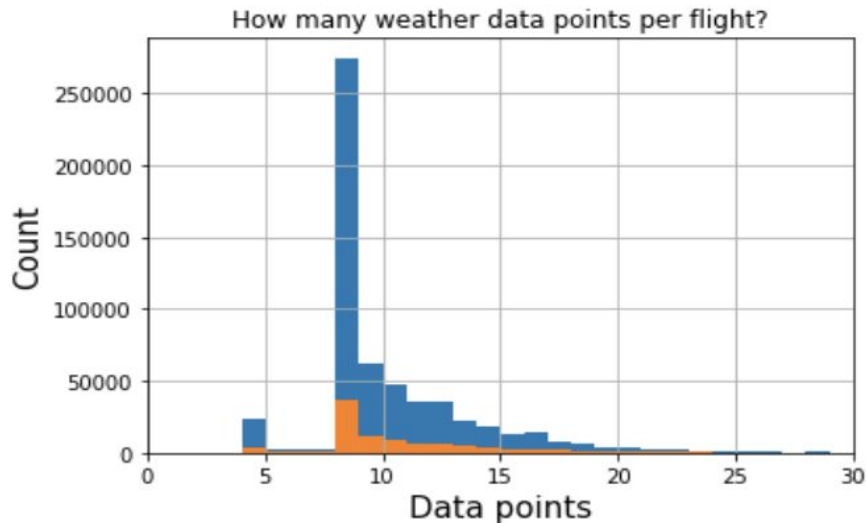
Table to match airport with timezone for conversion to UTC

<https://gist.github.com/mj1856/6d219c48697c550c2476#file-timezones-csv>

EDA



EDA - Weather

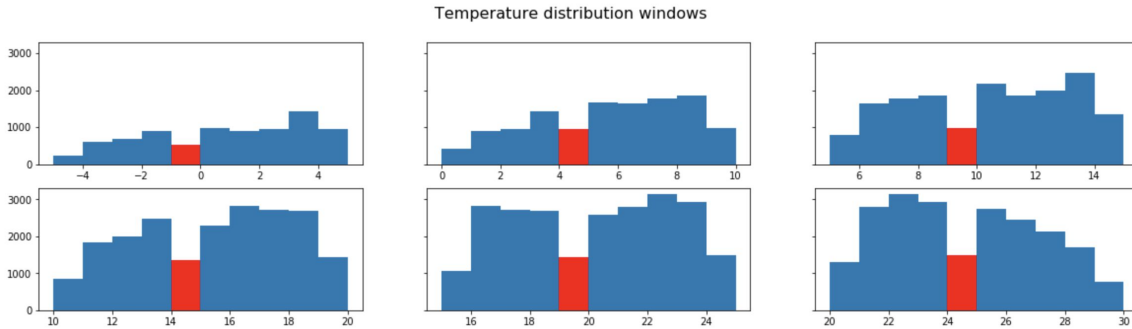


Delayed

On-time

EDA

A curiosity: Humans reading analog gauges?



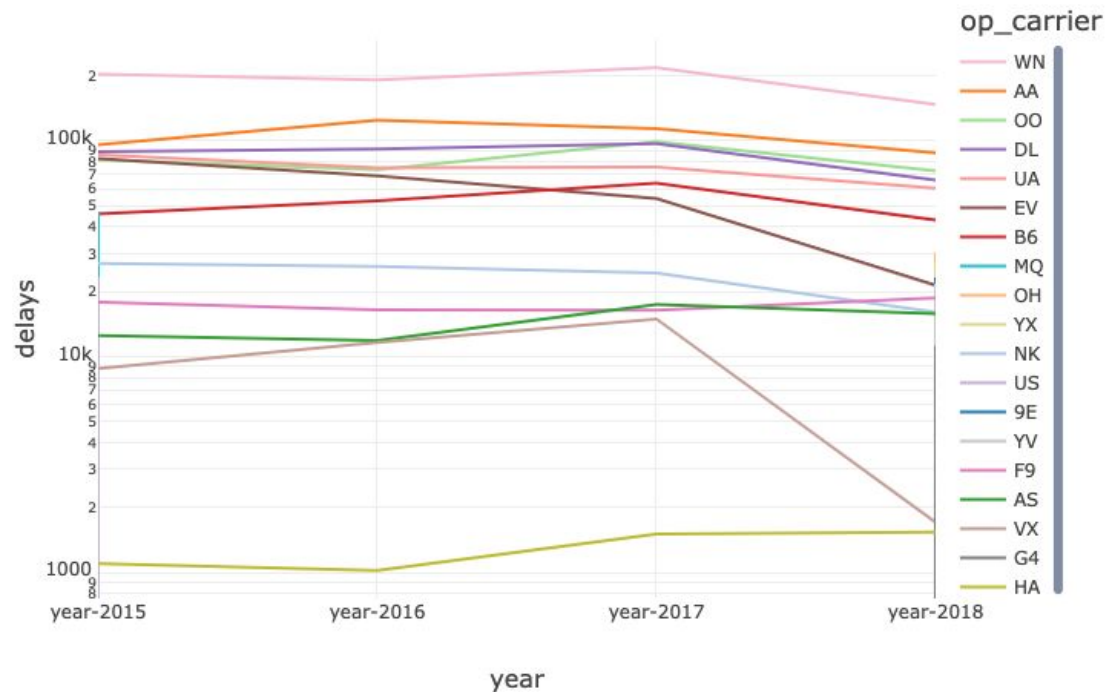
- Fewer 4s, 9s.
- But: Not enough 5s, 10s to explain the shortfall.



“I’ll just round that up”.

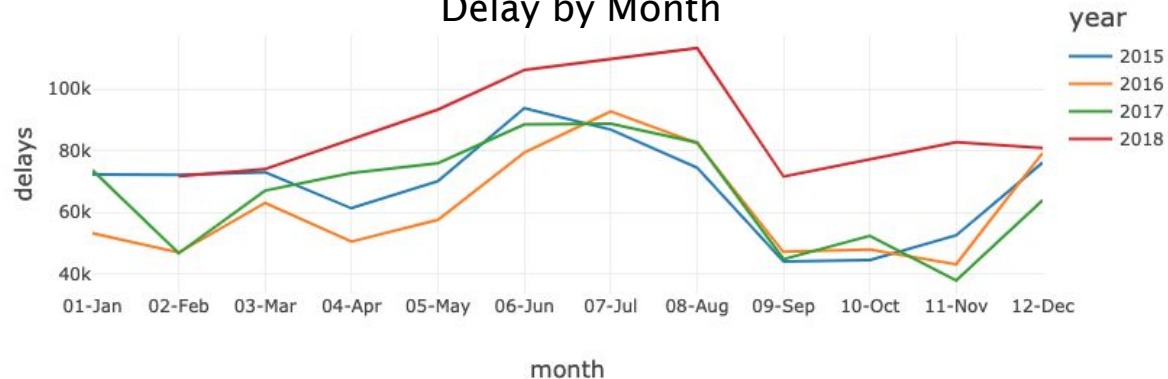
EDA

Yearwise Departure Delay by Carrier (log Scale)

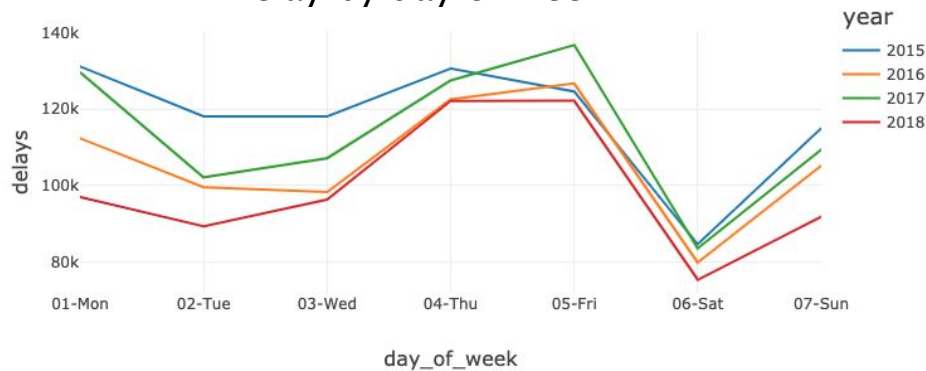


EDA

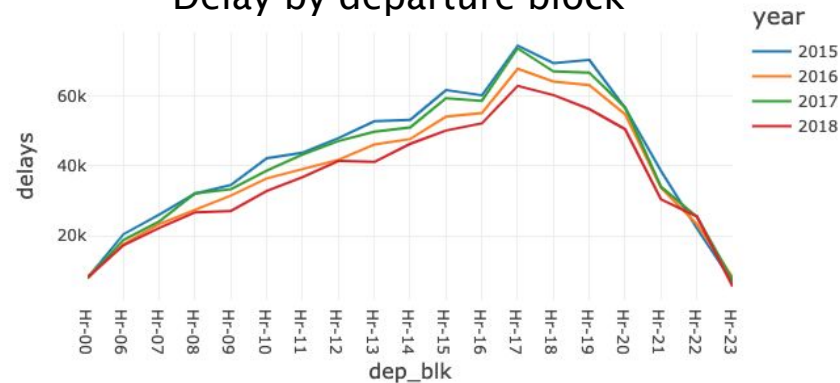
Delay by Month



Delay by day of week

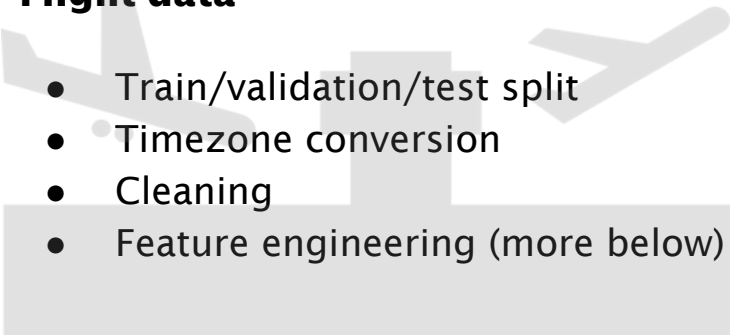


Delay by departure block

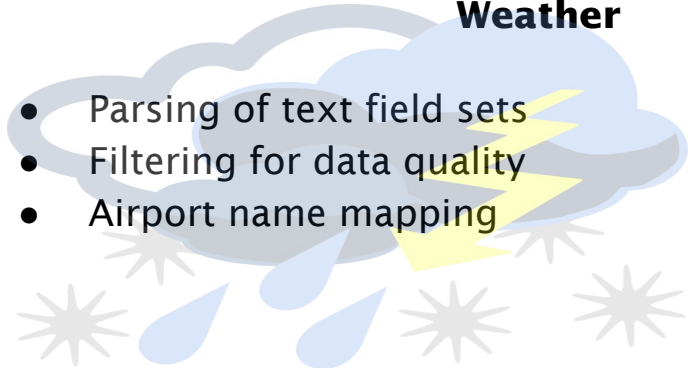


ETL

Flight data

- 
- Train/validation/test split
 - Timezone conversion
 - Cleaning
 - Feature engineering (more below)

Weather

- 
- Parsing of text field sets
 - Filtering for data quality
 - Airport name mapping

Join and Pivot

- Join each flight with 4 hours of weather for origin and destination
- Pivot weather into columns

Data Split Strategy

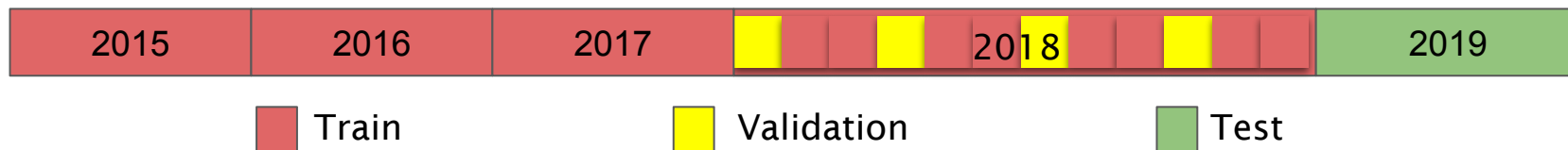


Table 2: Data Split Strategy

Data Set	Timing	Total months of data available	Num. of Records
Train	2015, 2016, 2017, 2018 (Feb/Mar/May/Jun/Aug/Sep/Nov/Dec)	44	20,716,594
Validation	2018 (Jan/Apr/Jul/Oct)	4	2,293,672
Test	2019	12	6,841,352

Feature Engineering

Flights



- Page rank of origin and destination
- Edge weight for origin/destination
- Average Daily Throughput information
 - Avg. flights per day
 - by carrier
 - by destination
- Average daily trips by aircraft
- Ripples (next slide)

Weather

-
- A decorative background graphic for the Weather section featuring a large, stylized blue cloud with a yellow lightning bolt striking it. Below the cloud are several grey, star-like shapes representing rain or snow.
- Key indicators for hourly window from 2 to 5 hours before departure time
 - Wind speed/angle
 - temperature/dewpoint
 - Visibility
 - pressure

Ripples of Delayed Aircraft

Delayed flights

<u>Tail num</u>	<u>Arr time</u>
4	2/20/2015 10:42
7	4/1/2017 15:27
2	5/20/2018
6	10/14/2017
3	9/5/2018

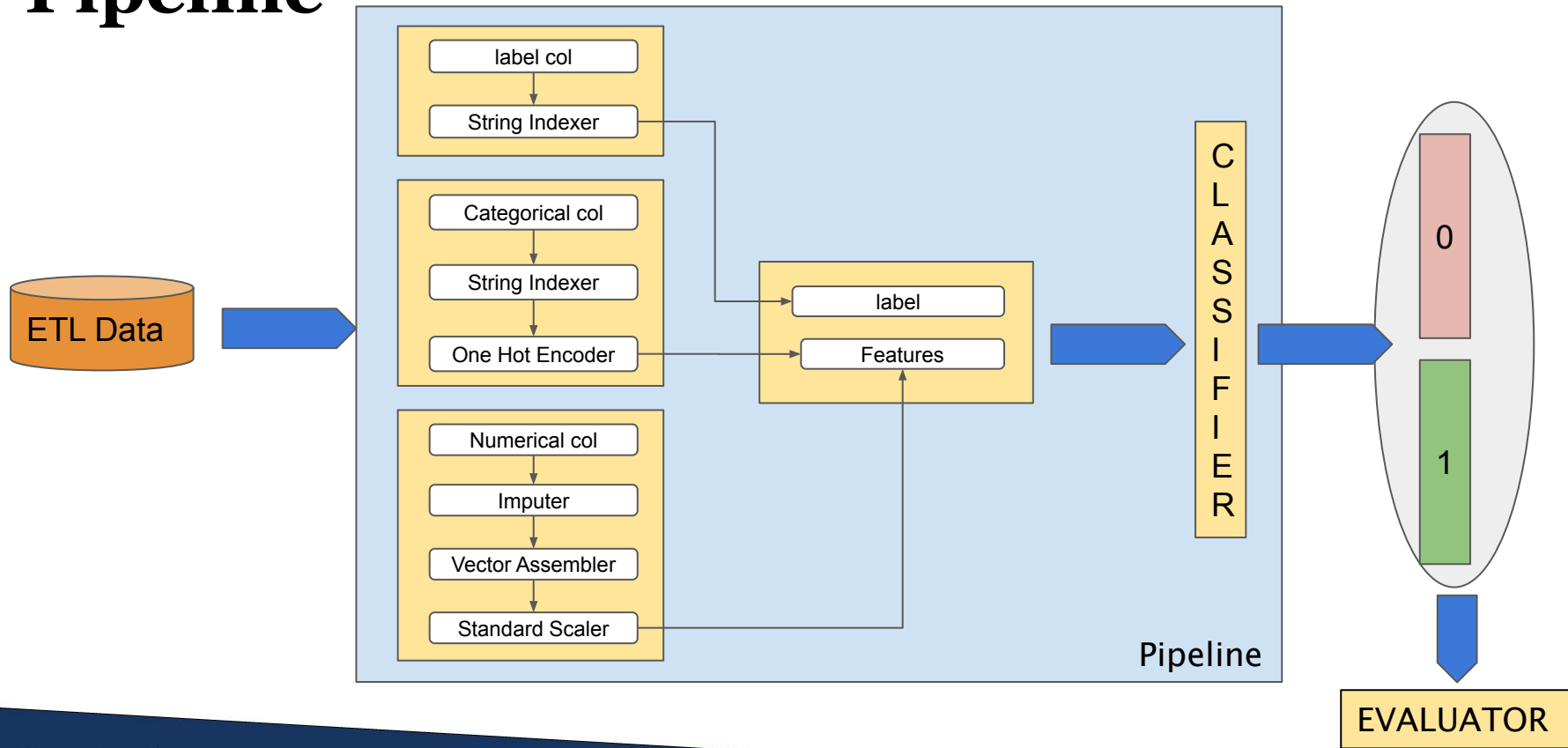
Delayed aircraft –
projected forward

<u>Tail num</u>	<u>Arr time</u>
7	4/1/2017 15:27
7	4/1/2017 16:27
7	4/1/2017 17:27
7	4/1/2017 18:27
7	4/1/2017 19:27
7	4/1/2017 20:27
7	4/1/2017 21:27

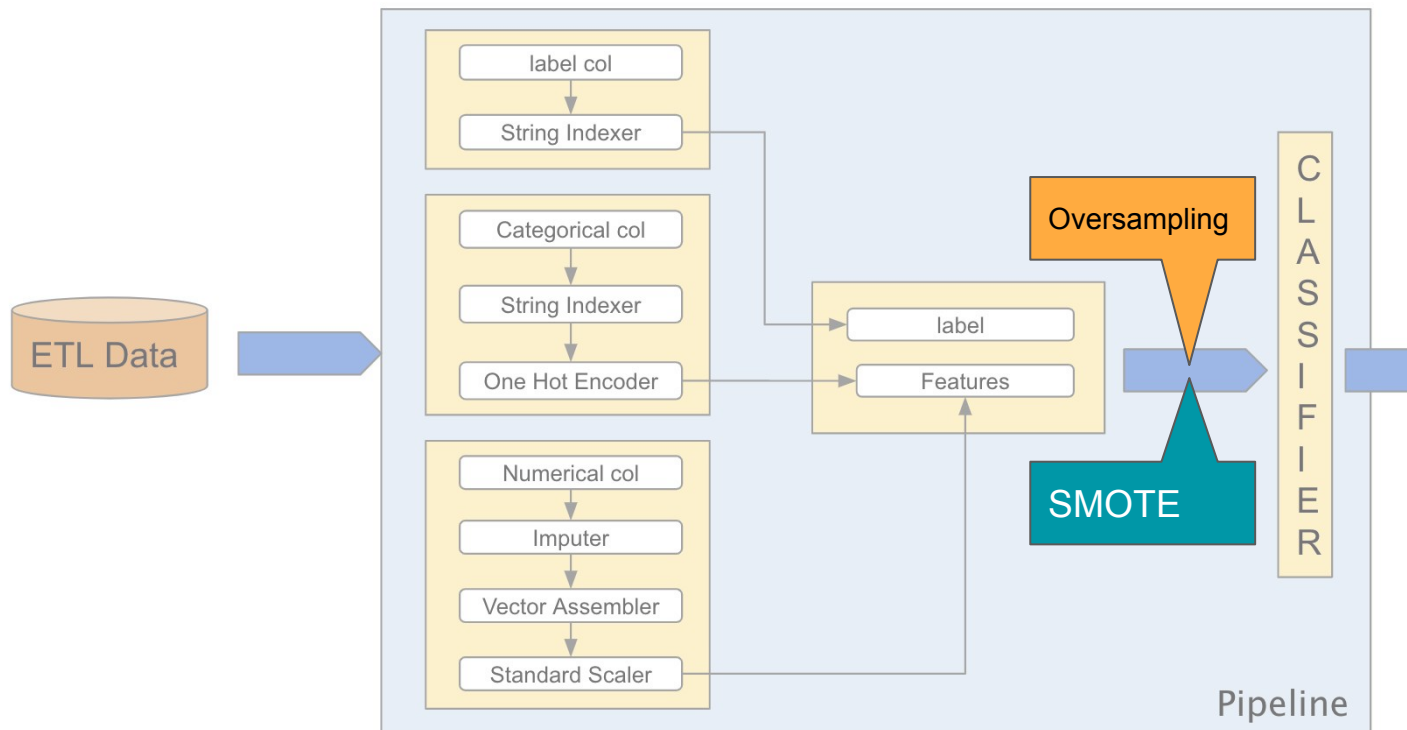
Delayed aircraft – matched
with future departures

<u>Tail num</u>	<u>Dep time</u>
56	2/20/2015 10:42
7	4/1/2017 19:45
10	5/20/2018
8	10/14/2017
9	9/5/2018

Pipeline



SMOTE and oversampling



Algorithm Selection

Features	Metrics	Classifiers			
		Logistic (default)	SDT (default)	RF (depth:12)	GBT (default)
Basic Features	Confusion Matrix	[[1,946,149. 63] [318,193. 63]]	[[1,1946,212. 0.] [318,256. 0.]]	[[1,946,212. 0.] [318,256 0.]]	[[1,946,195. 17] [318,253 3]]
	Class 0 Precision	0.859	0.856	0.856	0.856
	Class 1 Precision	0.5	0	0	0.15
	Weighted Precision	0.808	0.739	0.738	0.738
	Weighted Recall	0.859	0.856	0.856	0.856
	Weighted F1 Score	0.795	0.794	0.794	0.794
Enhanced Features	Confusion Matrix	[[1,918,152. 4,298.] [331,971. 10,047.]]	[[1,918,864. 3,586.] [332,775. 9,243.]]	[[1,919,185. 3,265.] [332,923. 9,095.]]	[[1,918,444. 4,006.] [331,675. 10,343.]]
	Class 0 Precision	0.852	0.852	0.852	0.853
	Class 1 Precision	0.7	0.72	0.735	0.721
	Weighted Precision	0.829	0.832	0.834	0.833
	Weighted Recall	0.851	0.851	0.851	0.851
	Weighted F1 Score	0.789	0.788	0.788	0.789

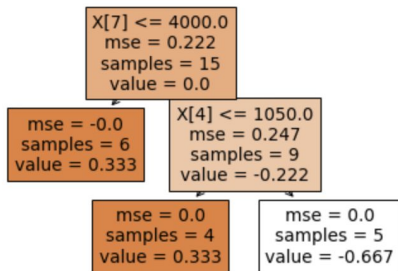
Algorithm Implementation

Gradient Boosting Tree Ensemble:
Successively correcting errors with
decision trees

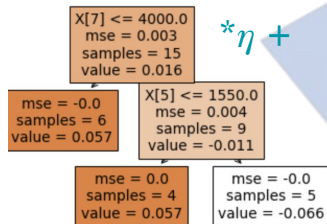
$$F_0 = \text{MLE}$$

	predictions ▲	residuals ▲
1	0.6667	0.3333
2	0.6667	-0.6667
3	0.6667	0.3333
4	0.6667	0.3333
5	0.6667	0.3333

$+\eta^*$

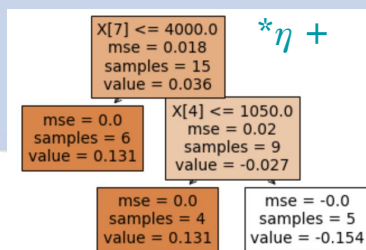


$*\eta +$



	predictions ▲	residuals ▲
1	0.9434	0.0566
2	0.0659	-0.0659
3	0.9434	0.0566
4	0.9434	0.0566
5	0.9434	0.0566

$*\eta +$



	predictions ▲	residuals ▲
1	0.8691	0.1309
2	0.1536	-0.1536
3	0.8691	0.1309
4	0.8691	0.1309
5	0.8691	0.1309

Fine Tuning

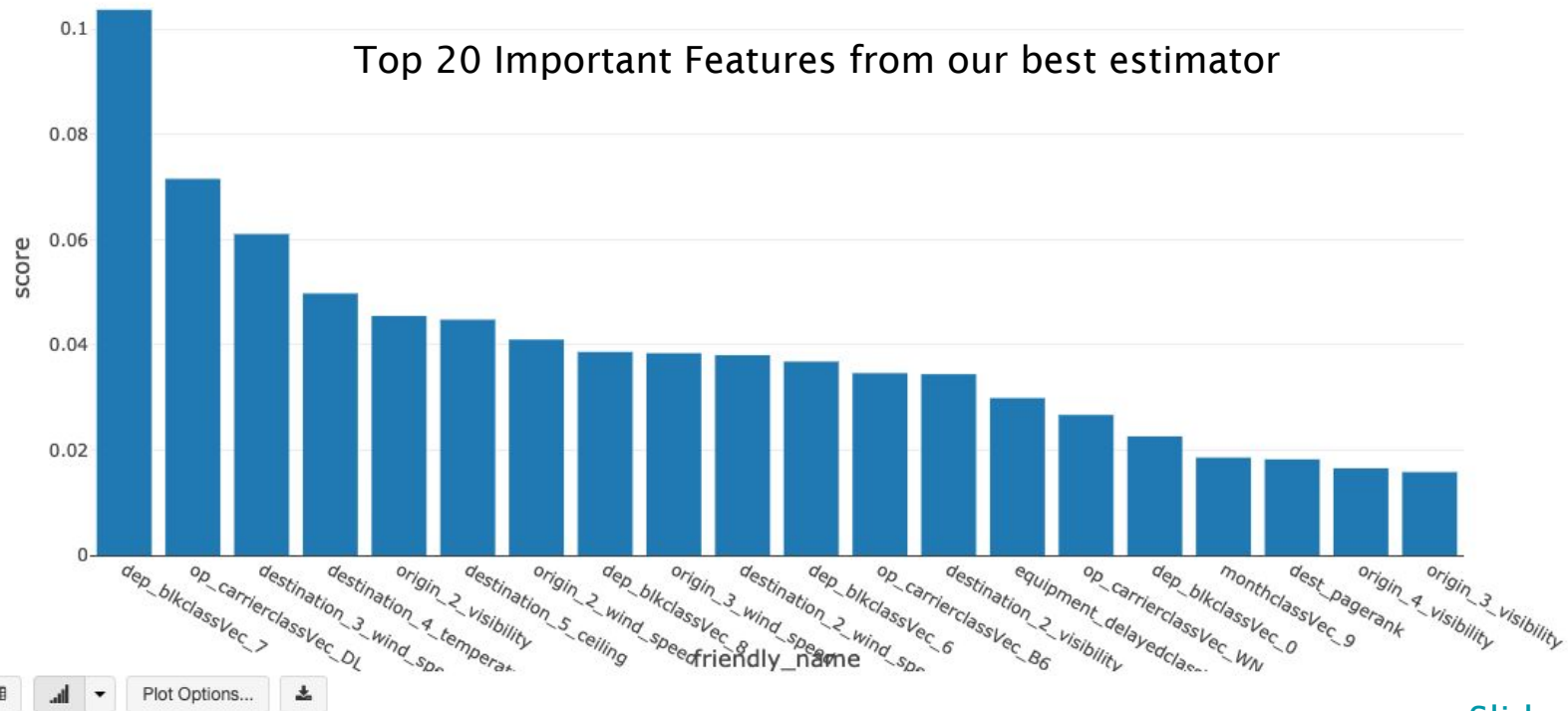
- Use 1% of training data and 1% of val data maintaining the ratio of the labels
- Use Random Search
- Use train validation split with 80/20 ratio
- Use validation tolerance to enable early stopping
- MLFlow to log metrics
- Evaluate on val data after each estimator run using weighted precision as the metrics
- Hyper parameters tuned
 - Learning rate = 0.064
 - Tree depth = 4
 - Number of iterations = 14
 - Max Bins = 27

	Parameters	Metrics	
Start Time	stepSize	val_weightedPrecision	weightedPrecision
2020-12-04 14:59:35	0.136	-	0.822
2020-12-04 14:59:35	0.064	-	0.83
2020-12-04 14:59:34	0.166	-	0.825
2020-12-04 14:59:33	0.1	-	0.827
2020-12-04 14:59:32	0.085	-	0.826
2020-12-04 14:59:31	0.137	-	0.823
✓ 2020-12-04 14:19:00	-	0.835	-

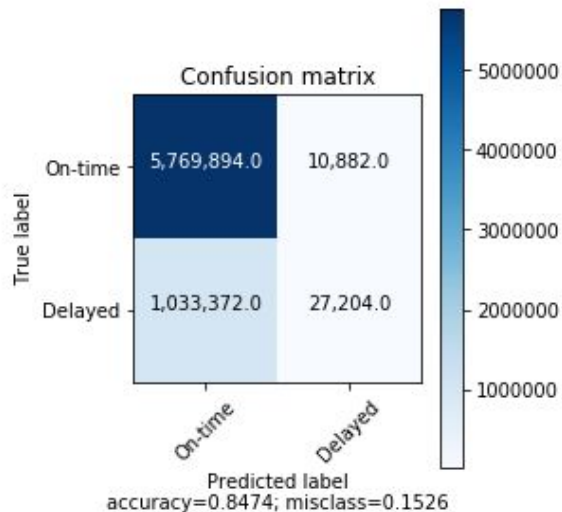
Results

Training Metrics	Testing Metrics
<pre>-----Training Metrics----- Test Area Under ROC: 0.6642795185272679 Test Area Under PR: 0.2761523705320385 Accuracy = 0.853929 Test Error = 0.146071</pre>	<pre>-----Testing Metrics----- Test Area Under ROC: 0.6603749781039275 Test Area Under PR: 0.27798935421806714 Accuracy = 0.847361 Test Error = 0.152639</pre>

Results



Results



Summary Stats

Precision = 0.7142782124665231

Recall = 0.025650212714600367

F1 Score = 0.0495220550087288

Class 0 precision = 0.8481064829745008

Class 0 recall = 0.9981175537678678

Class 0 F1 Measure = 0.9170176005451984

Class 1 precision = 0.7142782124665231

Class 1 recall = 0.025650212714600367

Class 1 F1 Measure = 0.0495220550087288

Weighted recall = 0.8473614572090429

Weighted precision = 0.8273598452014015

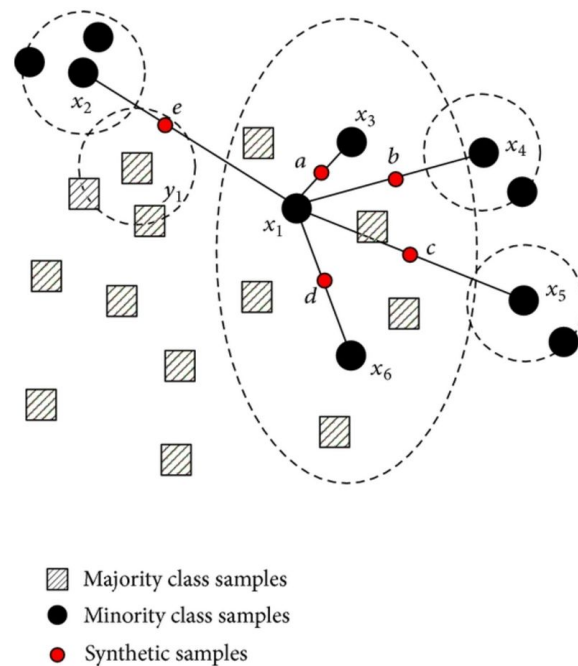
Weighted F(1) Score = 0.7825346861003801

Weighted F(0.5) Score = 0.7562227560061363

Weighted false positive rate = 0.8235936907265747

Handling Class Imbalance

- Delayed to On time flights is at 1: 5 in training data and 1:7 in full dataset.
- Adjusting imbalance using oversampling of minority class.
 - minority class = majority class leads:
 - Huge jump in true positives(and false positives) and sharp decrease in false negatives.
Recall improved dramatically, Precision dropped significantly
- Adjusting imbalance using SMOTE (Synthetic Minority Oversampling Technique)
 - Use nearest neighbors using locality sensitive hashing
 - Features grouped by origin, destination, month
 - Categoricals chosen as-is between candidates
 - Numericals are interpolated
Precision dropped significantly



Additional Models

Departure time 5 to 7pm

- Class 1 precision 74.9%
- Class 1 recall 4%
- AUC Score 0.663

Winter holidays (Nov 22 - Jan 3)

- Class 1 precision 73.9%
- Class 1 recall 2.8%
- AUC Score 0.675

Conclusion

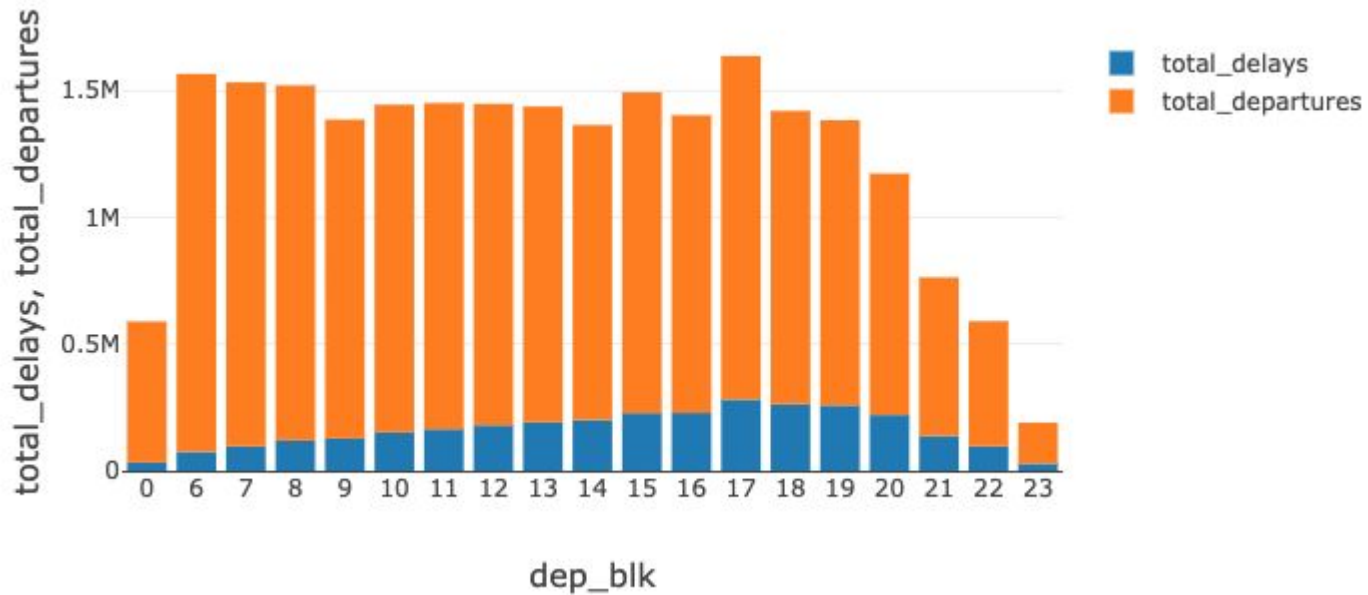
- Our goal was to maximize true positive and minimize false positive. Our delayed flight predictions have 71% precision.
- We are able to predict 99.8% of ontime flights correctly. However, our model is only able to predict 2.6% of the total delayed flights correctly.
- Overall model precision is at 83% and recall is 85%. Model accuracy is at 84.7%
- Model AUC of 0.66 indicates that model has discriminatory power to distinguish the two classes.

With these metrics, **it will not be practical to use this model to predict delay on a per flight basis**. Needs to be augmented with more features pertaining to delay type.

Can use it for driving business decisions in auxiliary services related to air travel.

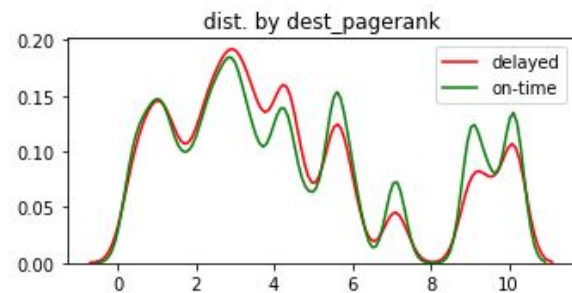
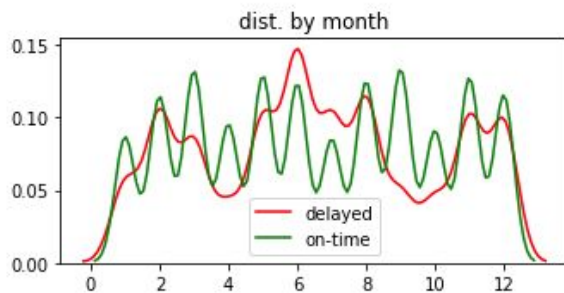
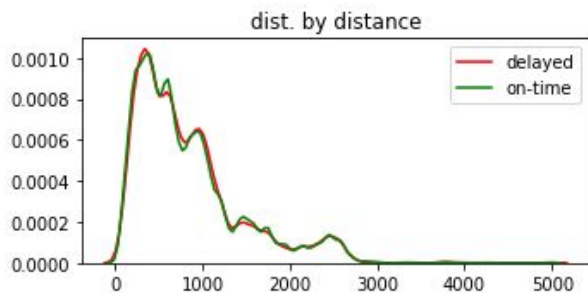
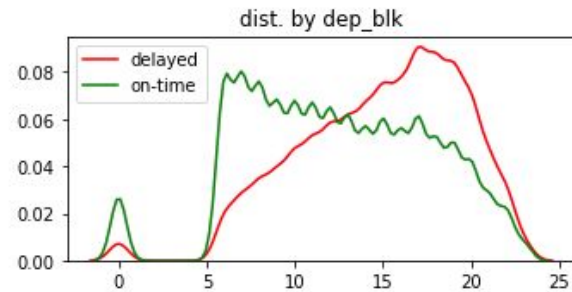
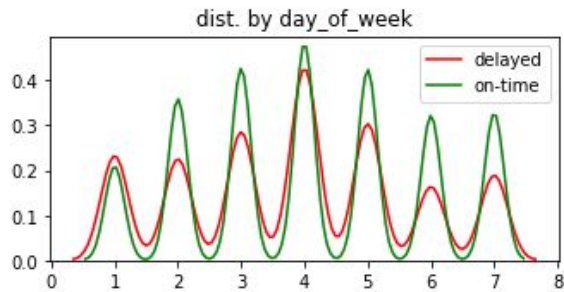
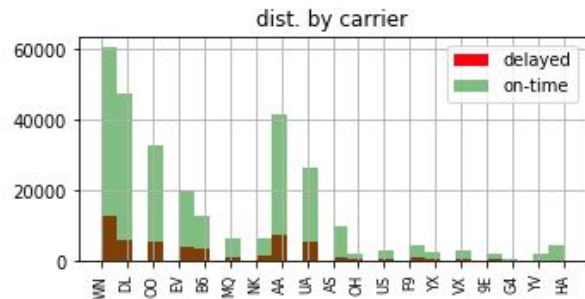
E.g. stocking up more food in nearby bistros, getting rooms ready for potential overnight delays etc.

Backup



Dist. of departures and delays by departure block

Distribution of Airport Features



Distribution of Weather Features

