Taming Turbulence with Machine Learning

Team 6-1

Dhyuti Ramadas

Juliana Gómez Consuegra

Jenna Sparks

Ray Cao

Rachel Gao

Outline

- Abstract & Project Description
- Exploratory Data Analysis & Feature Engineering
- Feature Refinement
- Modeling Results
- Discussion & Future Work
- Conclusion

Abstract
&
Project Description

Abstract & Project Description

Best model: Ensemble

Best pipeline: ML Flow Build

Primary evaluation metric: Fβ where β=0.5

standard_D4ads_v5 (1 master and 12 worker nodes) Databricks cluster

Cost Category	Description	Quantity/Hours	Rate/Unit Cost	Subtotal
Equipment Costs	Servers - Databricks	100	$ 6.40	$ 640.00
	Software Licenses	1	$ 5,000.00	$ 5,000.00
	GPU Units	1	$ 700.00	$ 700.00

Data

Flights

Airports

Weather
&
Weather Stations

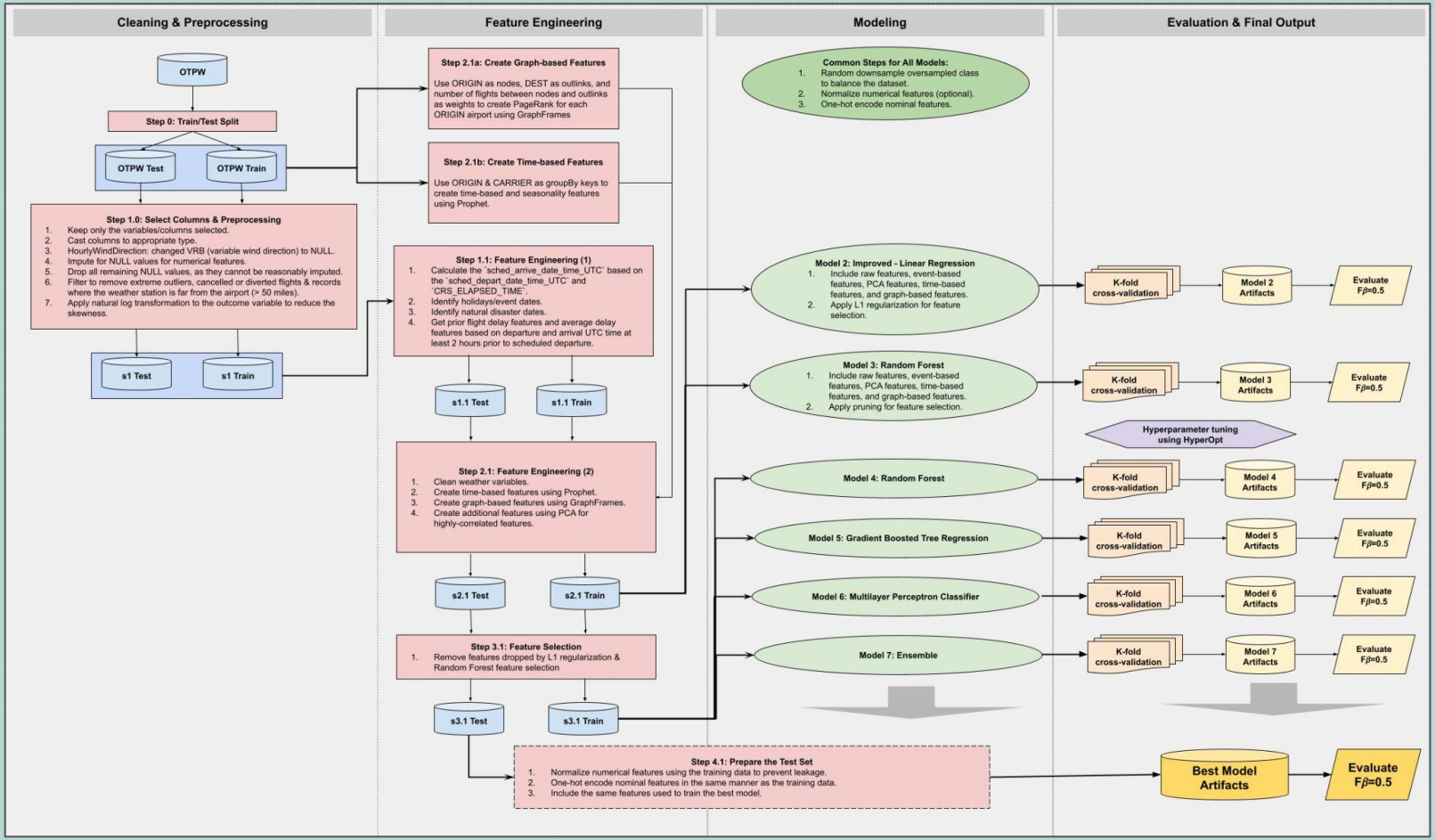
US On Time Performance Weather (OTPW)

- Train: 2015 - 2018
- Test: 2019

Outcome Variable: Log Departure Delay in Minutes

Class Imbalance: Delayed vs Canceled vs Diverted

Process Overview



Exploratory Data Analysis (EDA) & Feature Engineering

Overall Feature Engineering Informed by EDA

Seasonality Trend

- Quarter
- Month
- Day of Month
- Day of Week
- Holidays

Use Prophet to forecast delay based on Carrier & Origin Airports.

Airport Connectivity

- Origin Airports
- Destination Airports
- Number of Flights between Airports

Use GraphFrame to identify the most connected Origin Airports.

Aircrafts & Airports Delays

- Tail Number
- Two to Four Hours Before Scheduled Departure Time

Use Tail Number to identify prior aircraft delays.

Categorical Features

- Origin Airport
- Destination Airport
- Origin Airport Type
- Origin Airport Region
- Carrier
- Cloud Condition
- Disasters / Holidays / SuperBowl

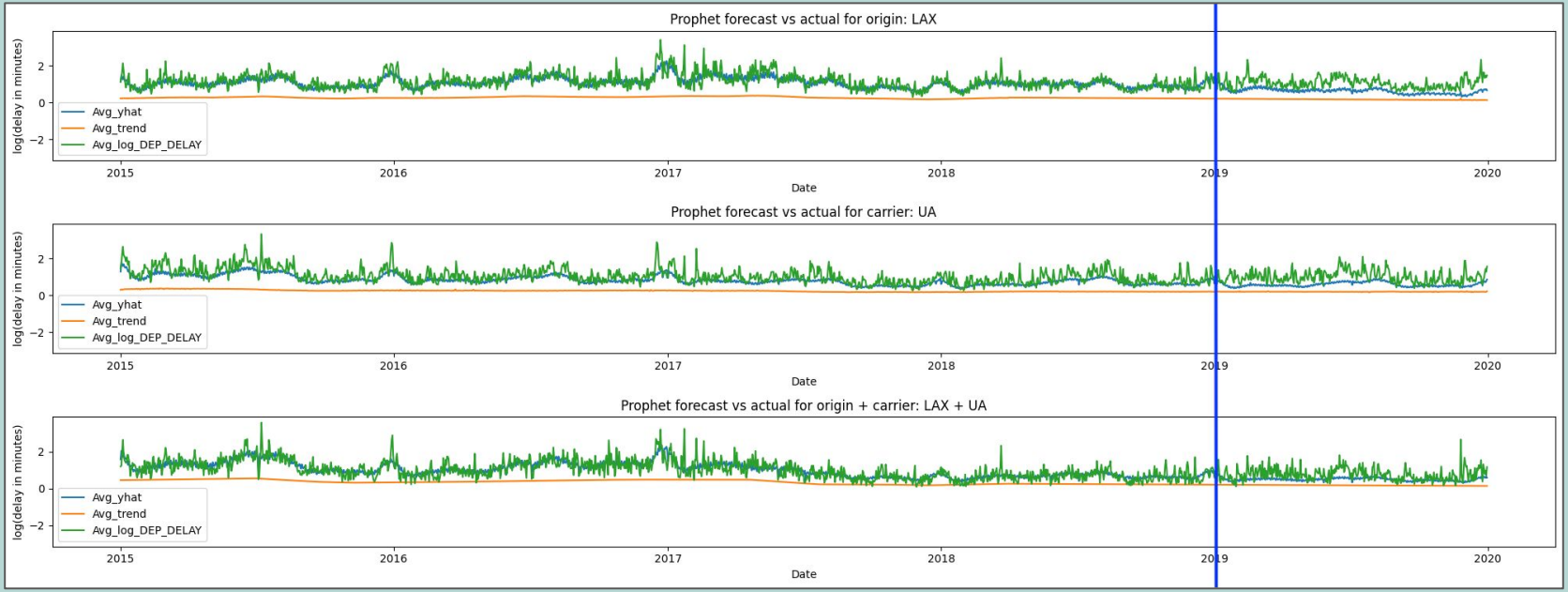
One-Hot Encode.

Highly Correlated Features

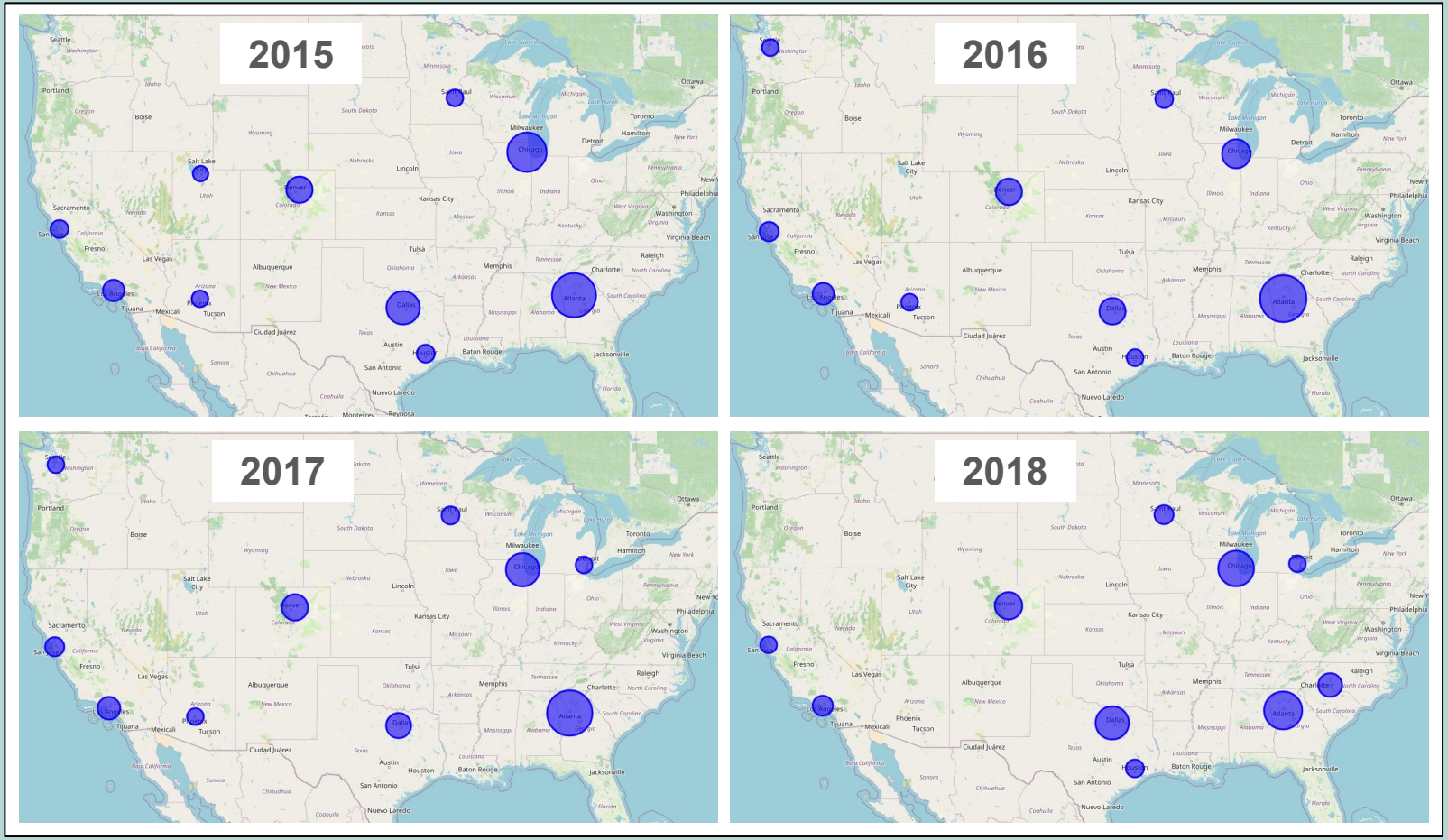
- Elevation / Station Pressure
- Altimeter / Sea Level pressure
- Scheduled Elapsed Time / Flight Distance
- Dew Point / Wind Chill / Wet Bulb / Dry Bulb

Apply Principal Components Analysis (PCA).

Seasonality Trend



Airport Connectivity



Aircrafts & Airports Delays

last delay

Icon by Hilmy Abiyyu Asad on freeicons.io

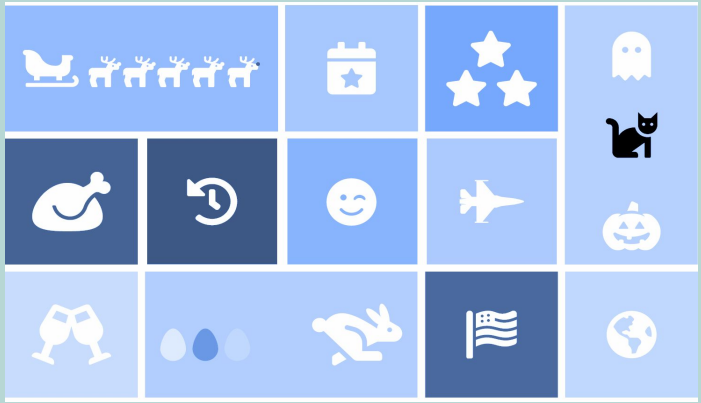
incoming flight delay ratio

Icon by [icon king] on freeicons.io

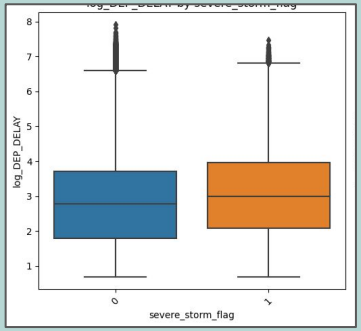
log average delay

Icon by Aficons on freeicons.io

Events and Natural Disasters

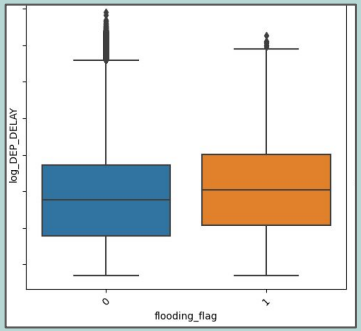


Severe storm



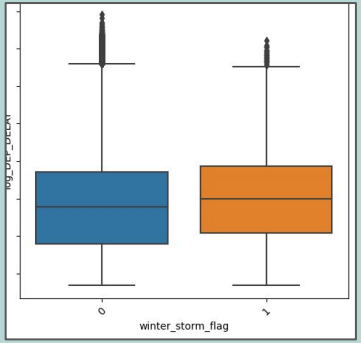
Icon by Ahbib Farrih on freeicons.io

Flooding



Icon by Anu Rocks on freeicons.io

Winter storm



Icon by Raj Dev on freeicons.io

Principal Components Analysis (PCA)

Elevation / Station pressure

Pearson R = -1.0

Icon by Raj Dev on freeicons.io

Altimeter / Sea level pressure

Pearson R = 0.9

Icon by BECRIS on freeicons.io

Elapsed Time / Distance between airports

Pearson R = 1.0

Icon by Gan Khoon on freeicons.io

Dew point / Wind chill / Wet bulb / Dry bulb

Pearson R = 0.8-0.9

Icon by ColourCreatype on freeicons.io

Feature Refinement

Feature Refinement

Keep

Variable Name	Data Type	Variable	Source	Decision Tree Importance	L1 Coefficient	Average
DEP_DELAY	float	Target	Flight	-	-	-
log_DEP_DELAY	float	Target	Flight	-	-	-
DEP_DEL15	boolean	Target	Flight	-	-	-
last_delay	float	Engineered Feature	Flight	67.19%	2.05%	34.62%
ORIGIN	string	Raw Feature	Flight	<0.1%	32.32%	16.16%
DEST	string	Raw Feature	Flight	1.36%	30.43%	15.90%
log_average_delay	float	Engineered Feature	Flight	26.09%	1.70%	13.90%
OP_UNIQUE_CARRIER	string	Raw Feature	Flight	0.85%	7.59%	4.22%
origin_region	string	Raw Feature	Flight	<0.1%	6.02%	3.02%
MONTH	integer	Raw Feature	Flight	<0.1%	4.42%	2.21%
CloudHeightandDarkness	string	Engineered Feature	Weather	<0.1%	3.38%	1.69%
DAY_OF_MONTH	integer	Raw Feature	Flight	<0.1%	3.35%	1.67%
time_series_forecast	float	Engineered Feature	Flight	1.29%	<0.1%	0.65%
HourlyRelativeHumidity	float	Raw Feature	Weather	0.87%	0.41%	0.64%
CRS_ELAPSE_TIME	integer	Raw Feature	Flight	0.60%	0.47%	0.54%
DAY_OF_WEEK	integer	Raw Feature	Flight	<0.1%	0.90%	0.46%
incoming_flight_delay_ratio	float	Engineered Feature	Flight	0.12%	0.78%	0.45%
winter_storm_flag	boolean	Engineered Feature	Weather	<0.1%	0.67%	0.34%
DISTANCE	float	Raw Feature	Flight	0.56%	<0.1%	0.28%
severe_storm_flag	boolean	Engineered Feature	Weather	<0.1%	0.48%	0.24%
origin_type	string	Raw Feature	Flight	<0.1%	0.40%	0.20%
flooding_flag	boolean	Engineered Feature	Weather	<0.1%	0.34%	0.17%
pagerank	float	Engineered Feature	Flight	0.34%	<0.1%	0.17%
wildfire_flag	boolean	Engineered Feature	Weather	<0.1%	0.28%	0.14%
HourlyAltimeterSetting	float	Raw Feature	Weather	<0.1%	0.26%	0.13%
trend	float	Engineered Feature	Flight	0.26%	<0.1%	0.13%
tropical_cyclone_flag	boolean	Engineered Feature	Weather	<0.1%	0.22%	0.11%
drought_flag	boolean	Engineered Feature	Weather	<0.1%	0.17%	<0.1%
HourlyWindSpeed	float	Raw Feature	Weather	<0.1%	0.14%	<0.1%
event_flag	boolean	Engineered Feature	Weather	<0.1%	0.13%	<0.1%
pca_dew_windchill_wet_temp	float	Engineered Feature	Weather	<0.1%	<0.1%	<0.1%
pca_elevation_station_pressure	float	Engineered Feature	Weather	<0.1%	<0.1%	<0.1%

Discarded

Variable Name	Data Type	Source	Decision Tree Importance	L1 Coefficient	Average
sched_depart_date_time	timestamp	Flight	-	-	-
sched_depart_date_time_UTC	timestamp	Flight	-	-	-
two_hours_prior_depart_UTC	timestamp	Flight	-	-	-
four_hours_prior_depart_UTC	timestamp	Flight	-	-	-
sched_arrive_date_time_UTC	timestamp	Flight	-	-	-
YEAR	integer	Flight	-	-	-
QUARTER	integer	Flight	-	-	-
TAIL_NUM	string	Flight	-	-	-
OP_CARRIER_FL_NUM	string	Flight	-	-	-
origin_station_id	string	Weather	-	-	-
HourlySkyConditions	string	Weather	-	-	-
SkyDarkness	string	Weather	<0.1%	1.51%	0.76%
CloudHeight	string	Weather	<0.1%	1.39%	0.70%
HourlyDewPointTemperature	float	Weather	0.16%	0.10%	0.13%
HourlyDryBulbTemperature	float	Weather	<0.1%	<0.1%	<0.1%
HourlyWetBulbTemperature	float	Weather	<0.1%	<0.1%	<0.1%
WindChill	float	Weather	<0.1%	<0.1%	<0.1%
pca_time_distance	float	Flight	0.11%	<0.1%	<0.1%
HourlySeaLevelPressure	float	Weather	<0.1%	<0.1%	<0.1%
pca_altimeter_sea_level_pressure	float	Weather	<0.1%	<0.1%	<0.1%
ELEVATION	float	Weather	<0.1%	<0.1%	<0.1%
HourlyStationPressure	float	Weather	<0.1%	<0.1%	<0.1%
HourlyWindDirection	integer	Weather	<0.1%	<0.1%	<0.1%
freeze_flag	boolean	Weather	<0.1%	<0.1%	<0.1%

Modeling Results

Evaluation Framework

Regression not classification to start:

- Extra information gain from knowing the precise delay minutes
- Loss function:

MSE = $\sum_{i=1}^{D}(x_i - y_i)^2$

Validation Metric:

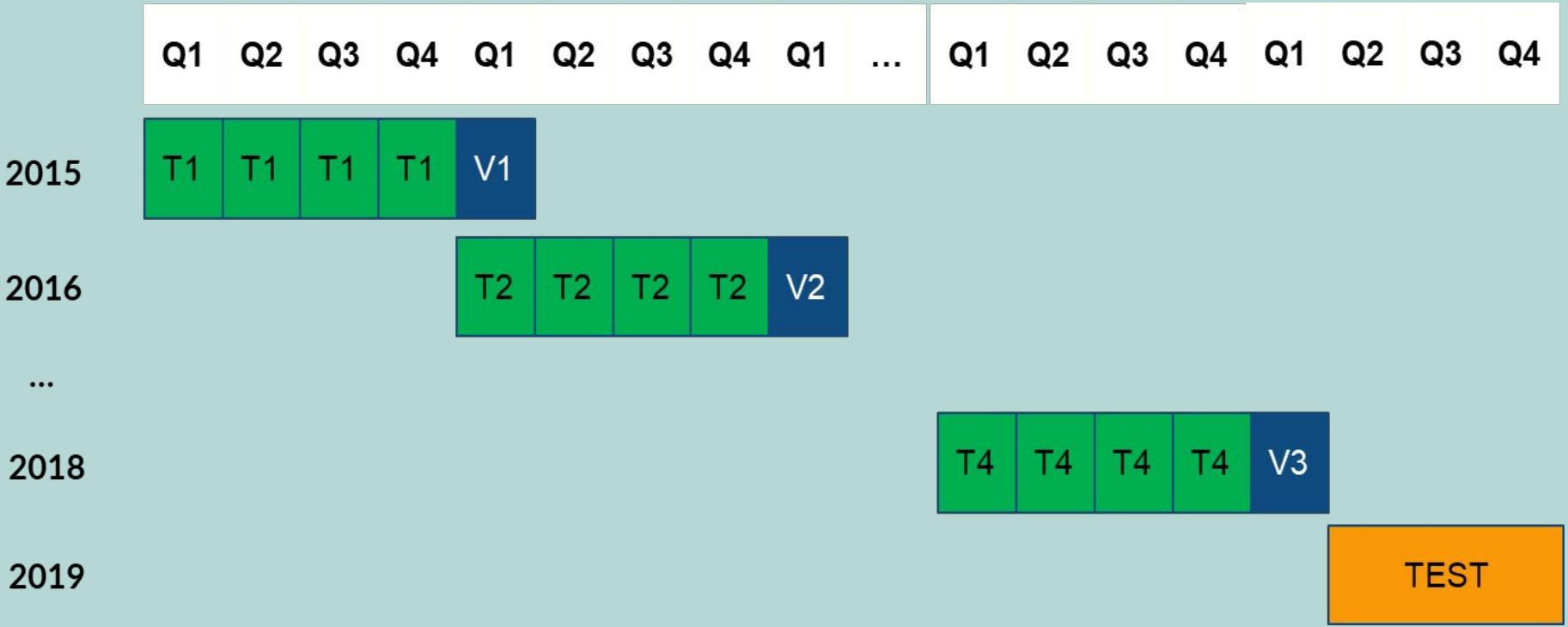
MAE = $\sum_{i=1}^{D}|x_i - y_i|$

Translate Regression results back to classification to end

- Final metric:
  - Fβ with β = 0.5

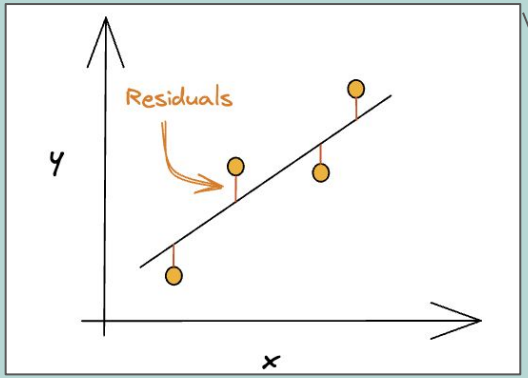
$$F_{beta} = \frac{(1+\beta^2)precision * recall}{\beta^2 * precision + recall}$$

Cross Validation

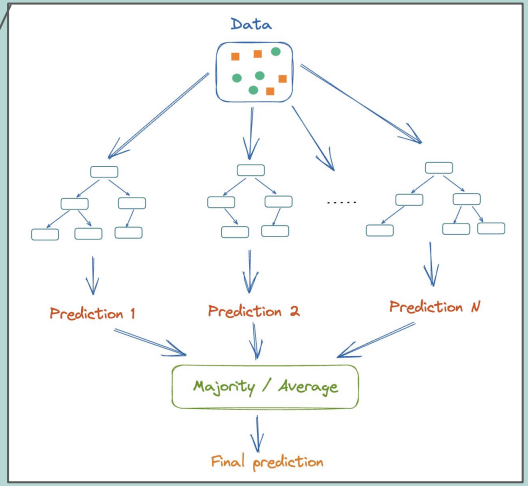


Modeling Algorithms

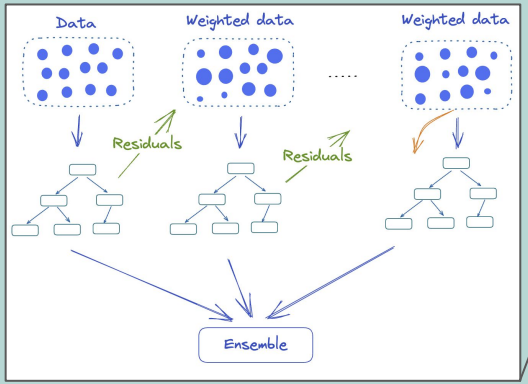
Linear regression + L2



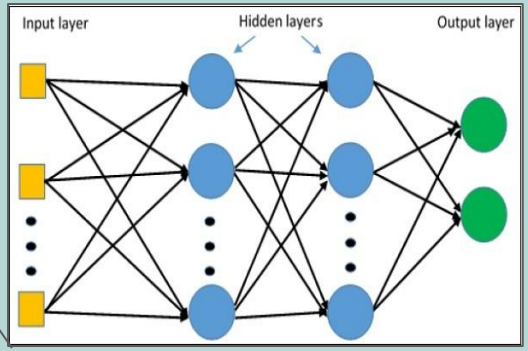
Random Forest



Ensemble



Gradient Boosted RF



Multilayer Perceptron

Source: https://illustrated-machine-learning.github.io/index.html#/machine-learning/ensemble#random-forest
https://www.tutorialspoint.com/tensorflow/tensorflow_multi_layer_perceptron_learning.htm

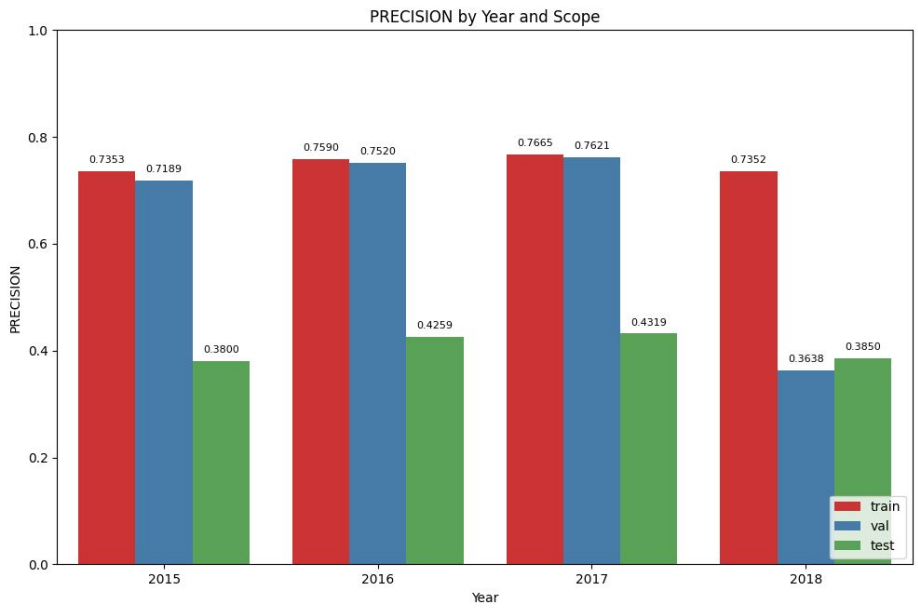
Train vs Val Result (For delayed flights, Average of models from each year)

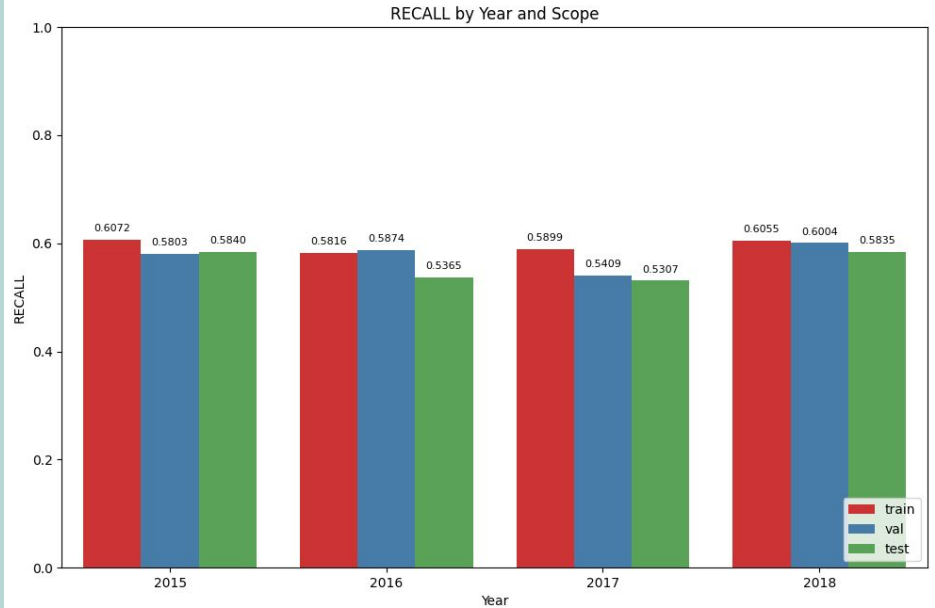
Model	Train			Validation		
	$F\beta$	Precision	Recall	$F\beta$	Precision	Recall
Baseline: Linear regression	0.52	0.76	0.23	0.35	0.63	0.13
Improved Baseline	0.68	0.87	0.37	0.67	0.87	0.35
Random Forest	0.69	0.89	0.37	0.67	0.90	0.34
Gradient Boosted Tree Regression	0.70	0.88	0.40	0.68	0.88	0.36
Multilayer Perceptron Classifier	0.71	0.76	0.55	0.70	0.76	0.52
Ensemble	0.71	0.76	0.60	0.65	0.65	0.58

Out of Sample Test Result (weighted average from cross val predictions)

Model	Regression			Classification		
	MAE	RMSE	MSE	$F\beta$	Precision	Recall
Baseline: Linear regression	20.8	49.5	2451.0	0.2455	0.2952	0.1468
Improved Baseline	17.7	46.0	2115.8	0.5223	0.5935	0.3529
Random Forest	16.9	45.3	2054.4	0.5552	0.6600	0.3396
Gradient Boosted Tree Regression	16.8	45.0	2026.2	0.5456	0.6257	0.3609
Multilayer Perceptron Classifier				0.4323	0.4128	0.5329
Ensemble				0.452	0.4362	0.529

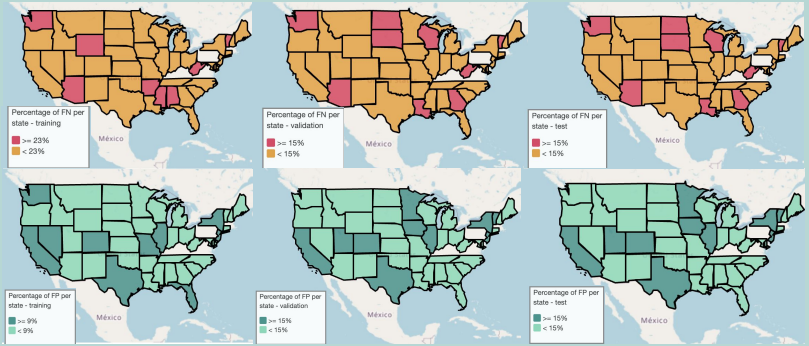
Inference From Ensemble Model

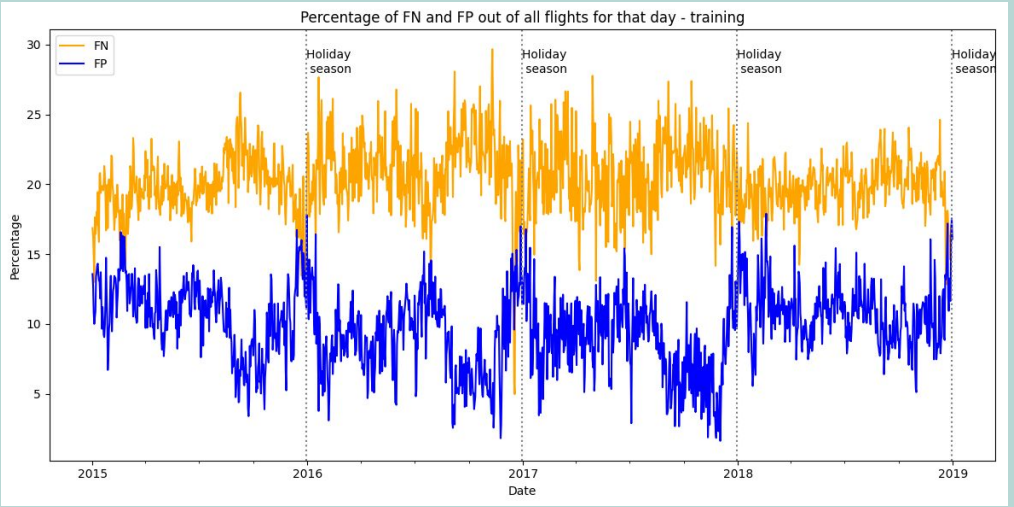


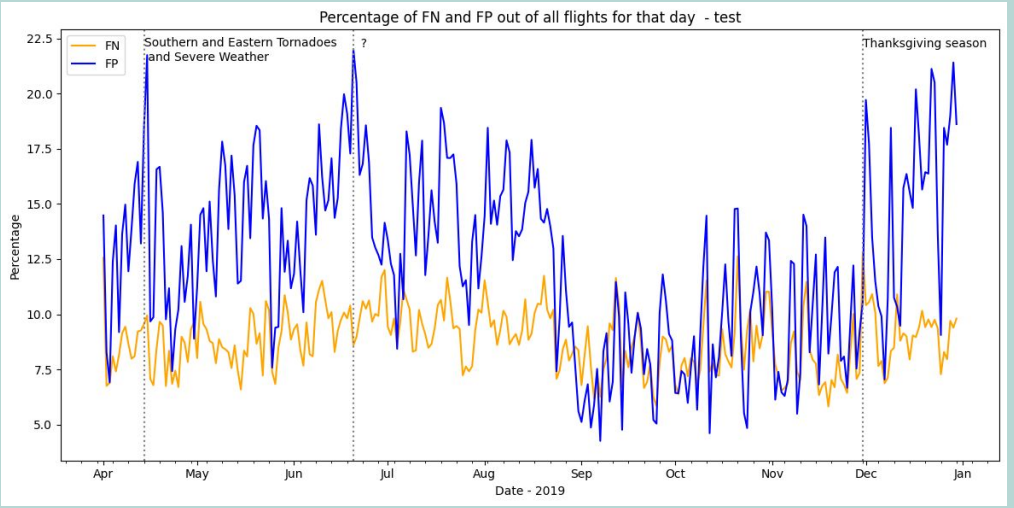


Gap Analysis
&
Future Work

Where is our model struggling?





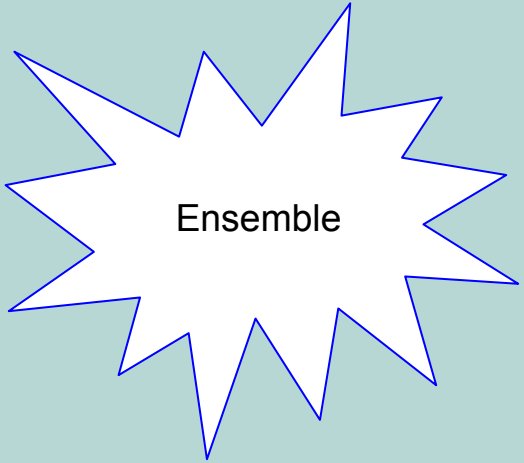


Conclusion

Conclusion

Flight delays

Can lead



To customer satisfaction

Icons by Gan Khoon Lay on freeicons.io

Icon by Wistudio on freeicons.io

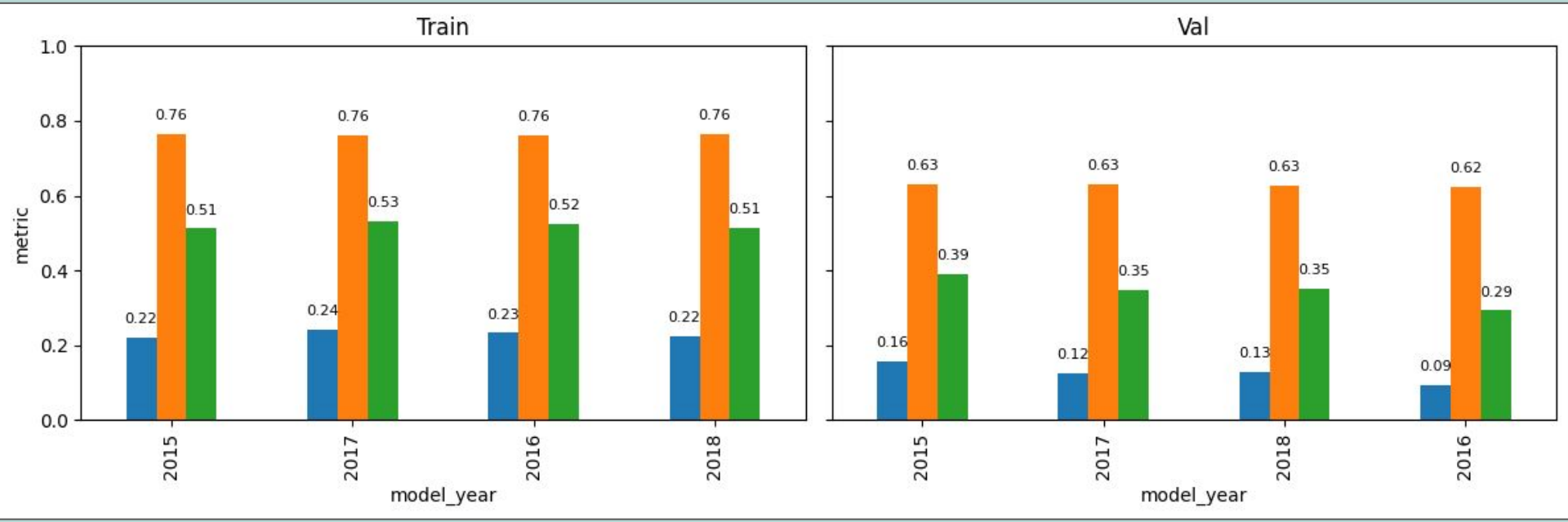
Thank you!

References

- Investigating The Costs And Economic Impact Of Flight Delays In The Aviation Industry And The The Potential Strategies For Reduction
- NOAA
- Flights Raw Data Source
- Weather Raw Data Source
- Airport Raw Data Source

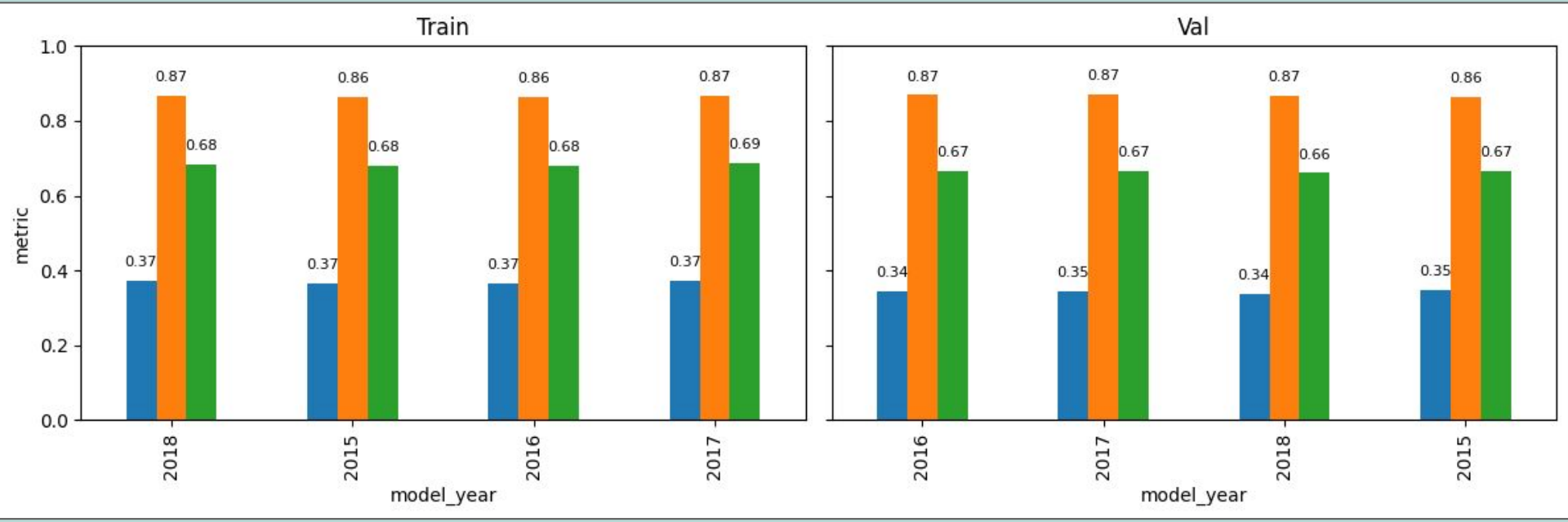
Baseline Model: Linear Regression

Average training time: 9 mins per year (at a smaller machine for POC)



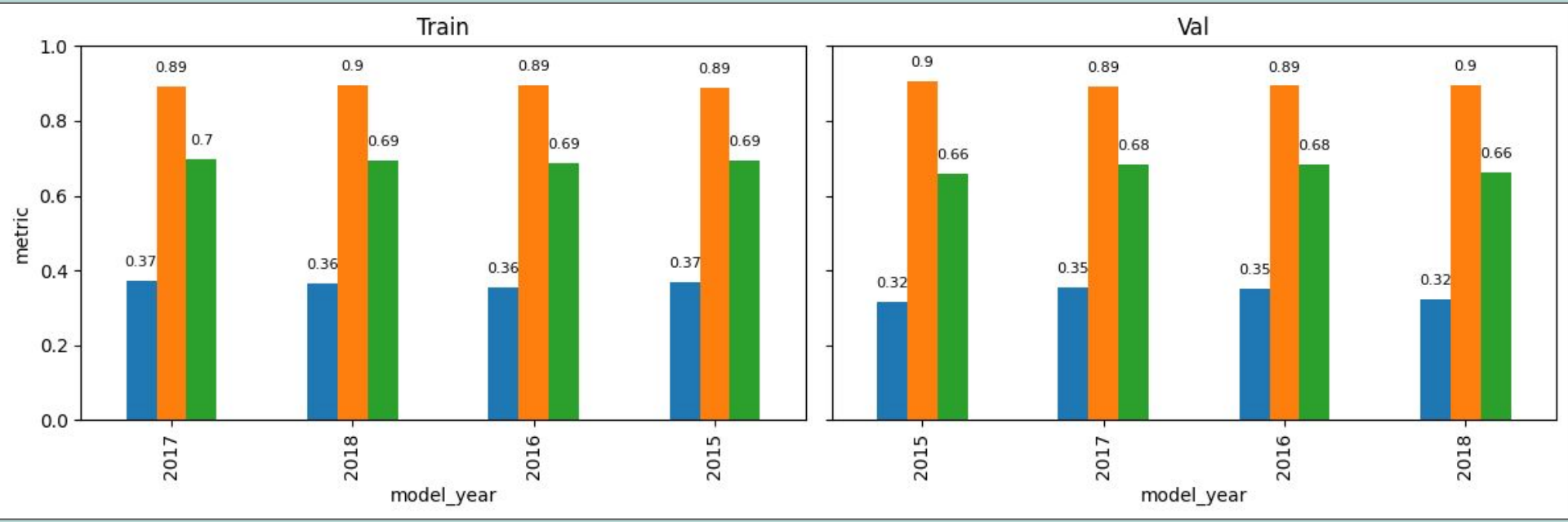
Improved Baseline: Linear Regression, L2 with Features

Average training time: 6 mins per year



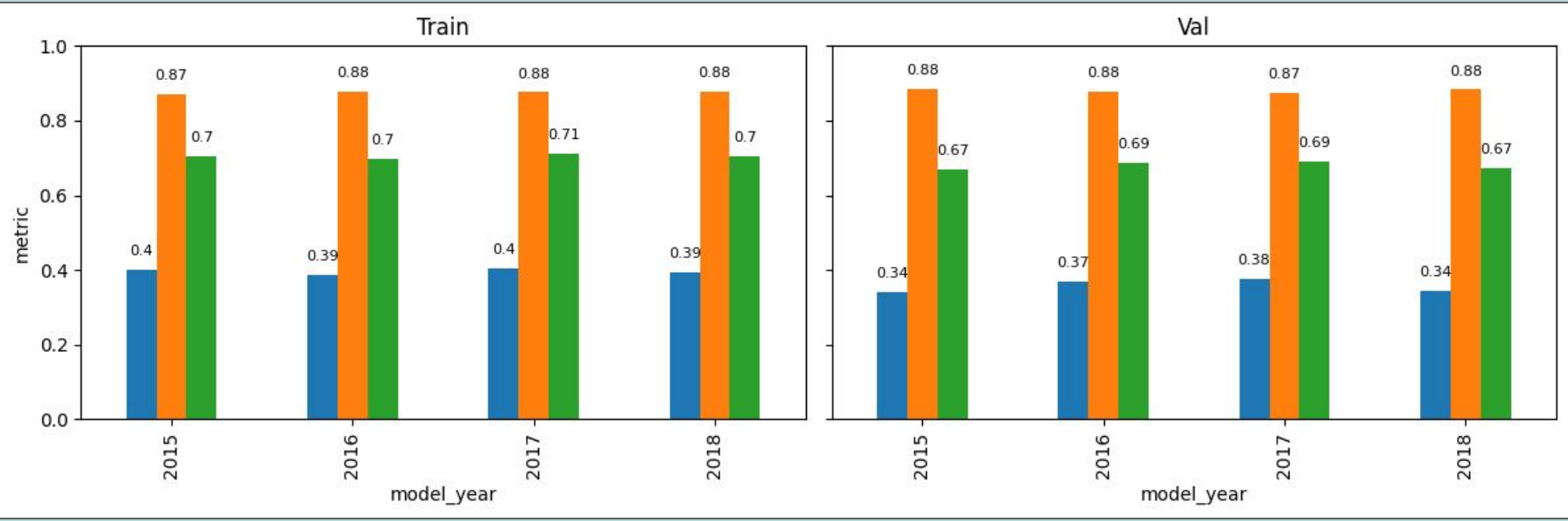
Random Forest

Average training time: 4 mins per year



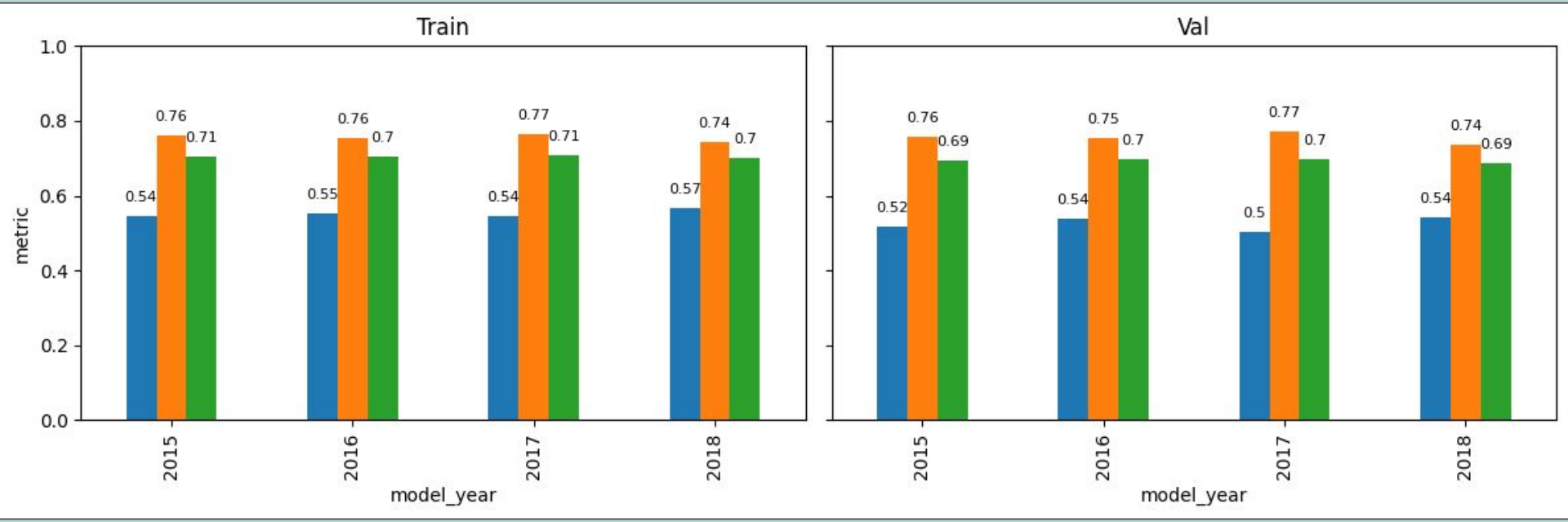
Gradient Boosted Random Forest

Average training time: 4 mins per year



Multilayer Perceptron Classifier

Average training time: 1.95 mins per year



Ensemble Model

Average training time: 1.95 mins per year