DataRobot vs. Open Source Forecasting Tools

Zachary Deane-Mayer

03 August, 2017

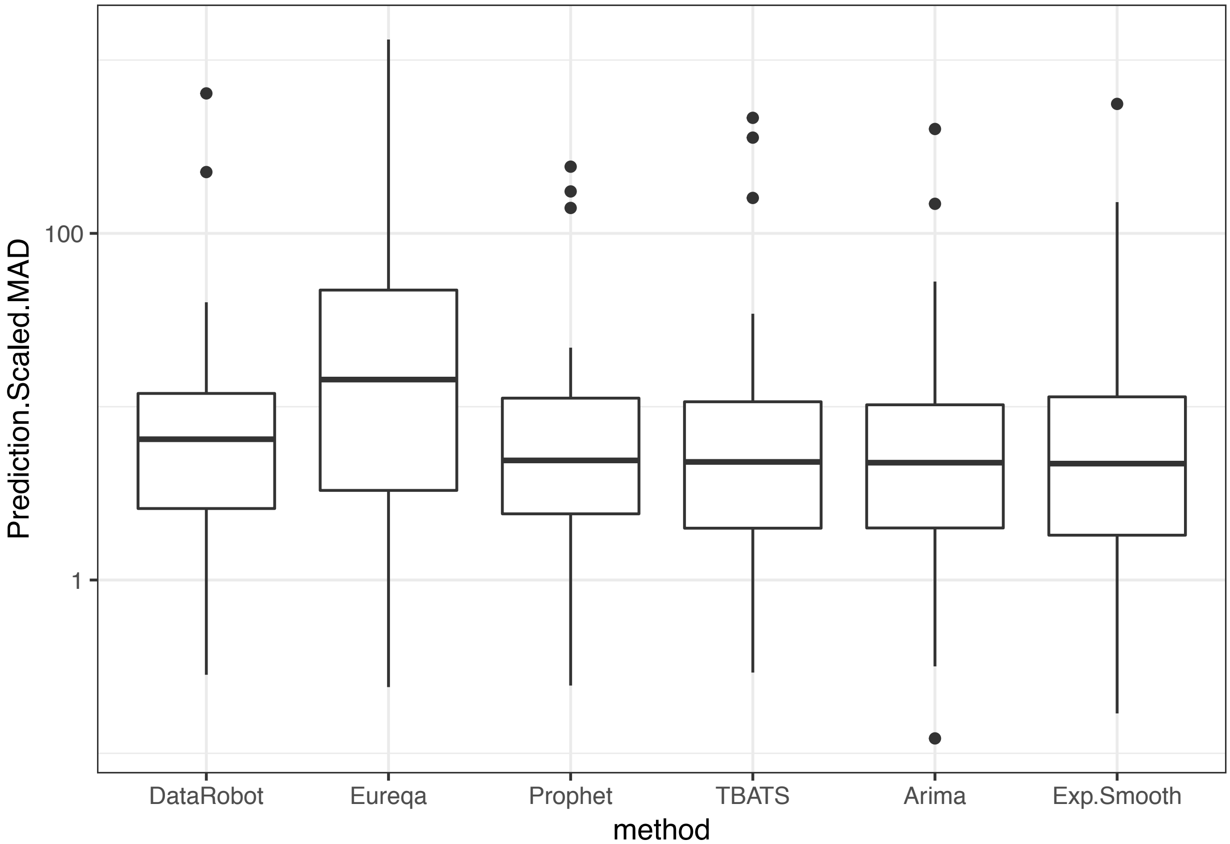
Methodology

We compared DataRobot's Out-of-Time Validation (OTV) models to Eureqa, R's Forecast package and Facebook's Prophet package on 52 datasets. We split these datasets into training sets and test sets, and ran each algorithm on the training set and calculated accuracy on the unseen test set. For DataRobot, we used the test set as a prediction set, so that it was not used by the autopilot for model training or selection. I used the metric "Normalized Gini Score" or "Gini Norm" to compare algorithms, which ranges from -1 (perfectly anti-predictive) to 0 (equivalent to random guessing) to 1 (perfectly predictive).

With time series problems, we see negative Gini scores more often than with traditional machine learning problems, as models can be fooled by randomness in the time series and end up extrapolating "trends" that do not exist.

We compared DataRobot and Eureqa to a total of 4 open source models: Prophet (from Facebook), auto.arima (from Forecast), ets (automated exponential smoothing from Forecast) and TBATS (a trigonometric function based model from Forecast). Note that all 4 of the open-source forecasting models minimize RMSE, so to keep the comparison fair, I used RMSE as the metric for all DataRobot projects. For each dataset, the DataRobot model with the best Gini Norm on the holdout set was used to make forecasts on the prediction set.

Results



On average, Eureqa is slightly more accurate than Datarobot (median Gini of 14.54 vs 6.49). DataRobot is slightly more accurate than Facebook's Prophet package (median Gini of 6.49 vs 4.91). TBATS is the best model from the Forecast package, but is typically worse that DataRobot or Prophet (median Gini of 4.8).

	method	min_smad	pct_25_smad	median_smad	mean_smad	sd_smad	pct_75_smad	max_smad
1	DataRobot	0.28	2.59	6.49	26.14	96.45	11.92	642.03
2	Eureqa	0.24	3.31	14.54	82.87	249.57	47.09	1313.07
3	Prophet	0.25	2.41	4.91	18.10	45.42	11.22	242.56
4	TBATS	0.29	1.99	4.80	27.07	84.80	10.68	463.99
5	Arima	0.12	2.00	4.75	18.64	60.50	10.25	400.21
6	Exp.Smooth	0.17	1.81	4.69	25.59	82.72	11.48	558.38

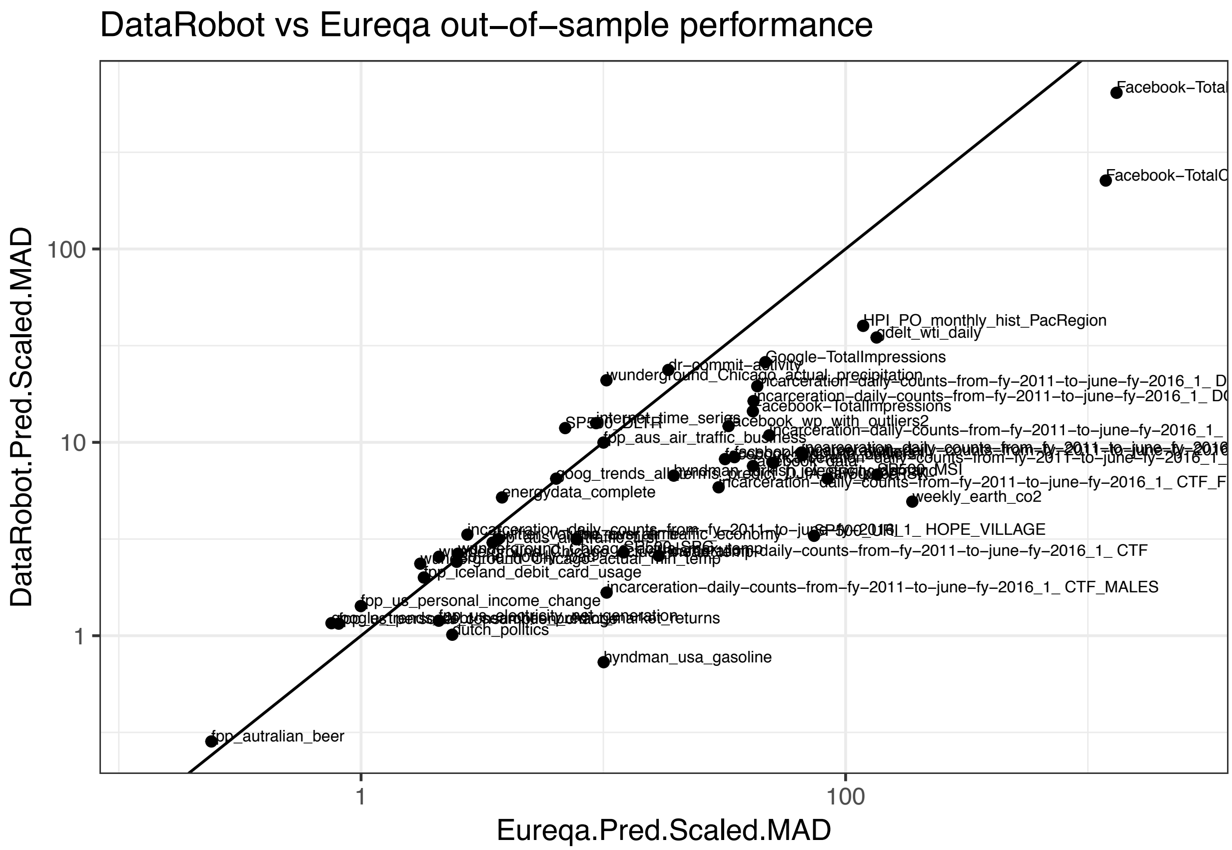
Table 1: Summary Results

Conclusion

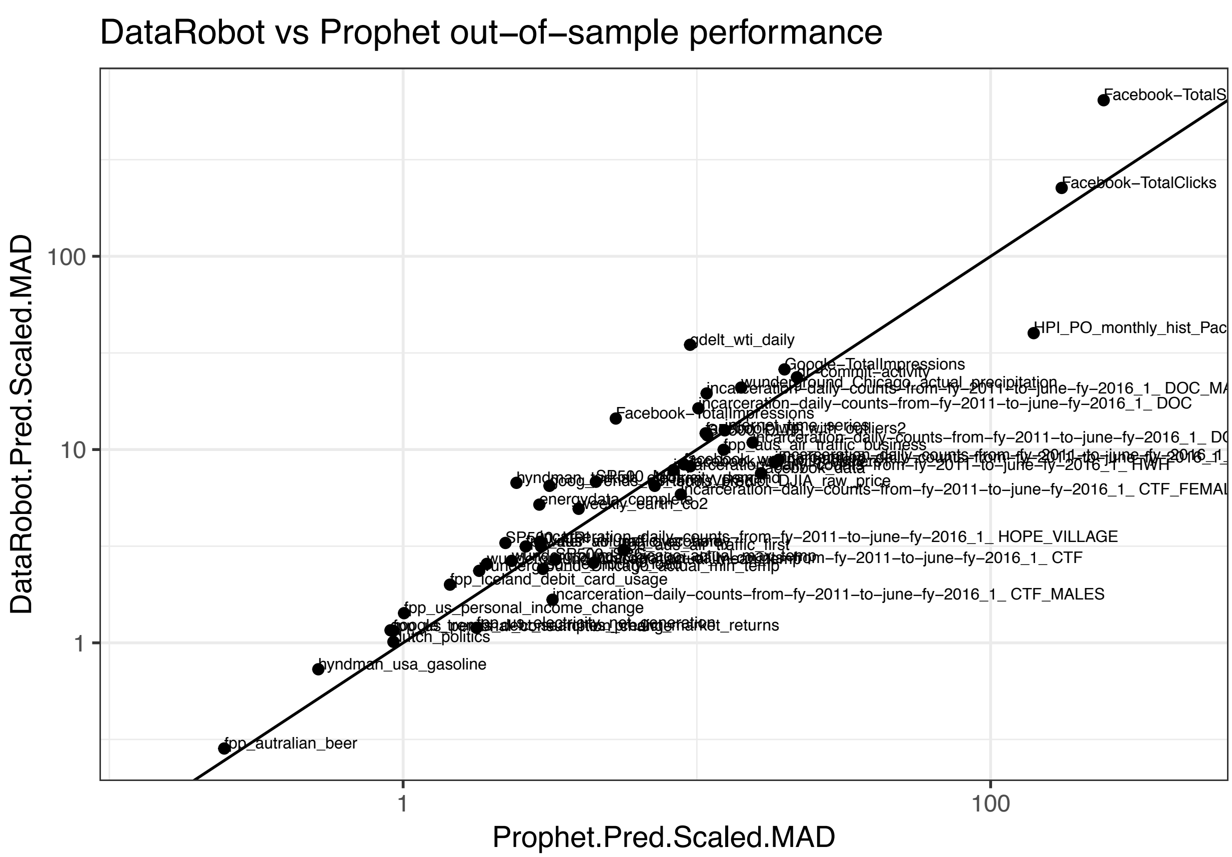
DataRobot's OTV models already perform well on a variety of time series problems, and are on average more accurate than all 4 of the open source forecasting models tried. DataRobot does especially well on datasets with complex seasonal patterns, e.g. electricity load, where the hour-of-day pattern may differ in a dramatic (and predictable) way between the winter and summer months. This sort of dynamic seasonality can be extremely difficult for traditional time series models to capture, but is modeled beautifully by DataRobot's "seasonal dummies + XGBoost" approach. DataRobot is also able to make use of covariates, while arima is the only open-source model with this capability.

However, there remains some room for improvement in DataRobot, which can be over-confident based on the holdout set when picking the model to use for forecasting. It might be beneficial to add some heuristics to DataRobot to cause it to prefer simpler models during model selection for time series models. This could help prevent DataRobot from picking complicated models that happen to get lucky in the holdout set but end up extrapolating the wrong trend. On random walk datasets like the S&P 500 stock market data, or the incarceration data, the correct forecast is usually a flat line from the last point (also known as a naive forecast), and the open source models tend to correctly predict this, while DataRobot tends to be overconfident in extrapolating trends.

We could gain additional accuracy in DataRobot, especially on more difficult problems, by adding Eurea and Arima blueprints to the autopilot.

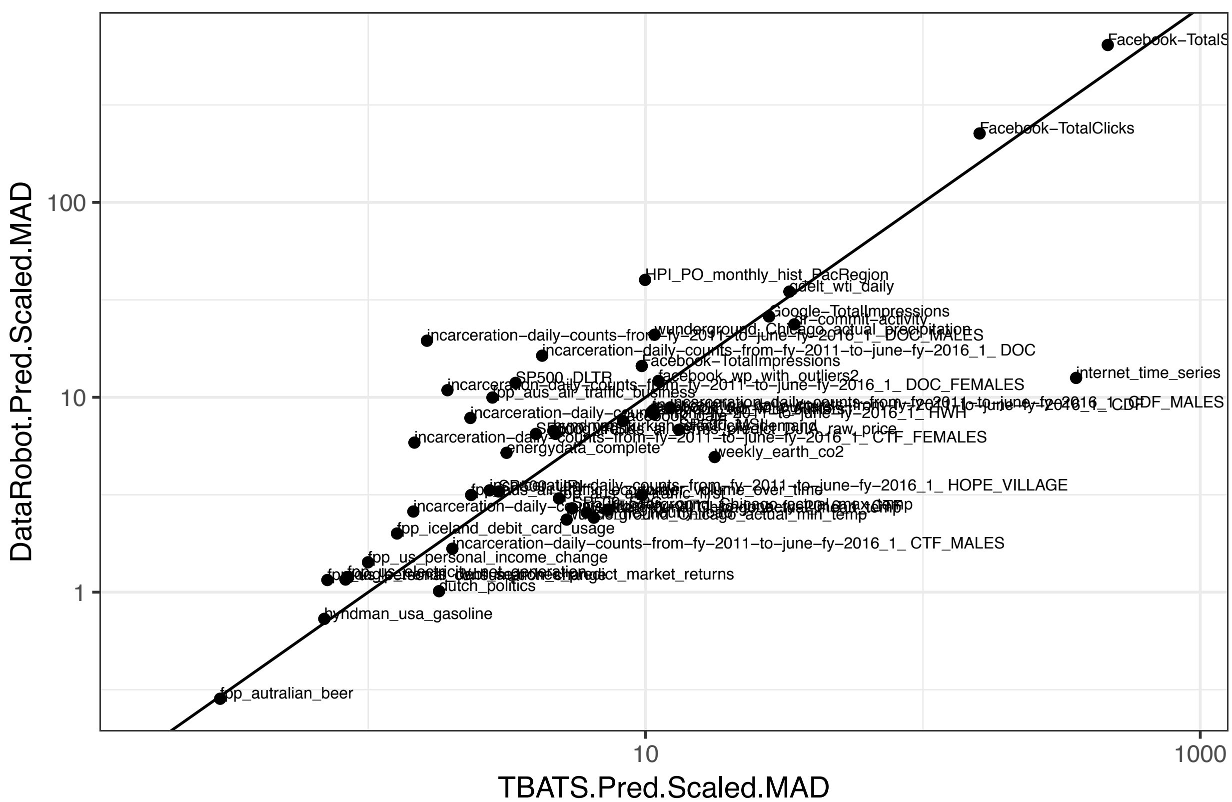


DataRobot vs Eureqa out−of−sample performance
DataRobot.Pred.Scaled.MAD
Eureqa.Pred.Scaled.MAD
1
10
100
Facebook−TotalClicks
HPI_PO_monthly_hist_PacRegion
gdelt_wti_daily
Google−TotalImpressions
dr−commit−activity
wunderground_Chicago_actual_precipitation
Facebook−TotalImpressions
internet_time_series
facebook_wp_with_outliers2
fpp_aus_air_traffic_business
energydata_complete
weekly_earth_co2
fpp_iceland_debit_card_usage
fpp_us_personal_income_change
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ HOPE_VILLAGE
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ CTF
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ CTF_MALES
fpp_us_electricity_net_generation
dutch_politics
hyndman_usa_gasoline
fpp_autralian_beer

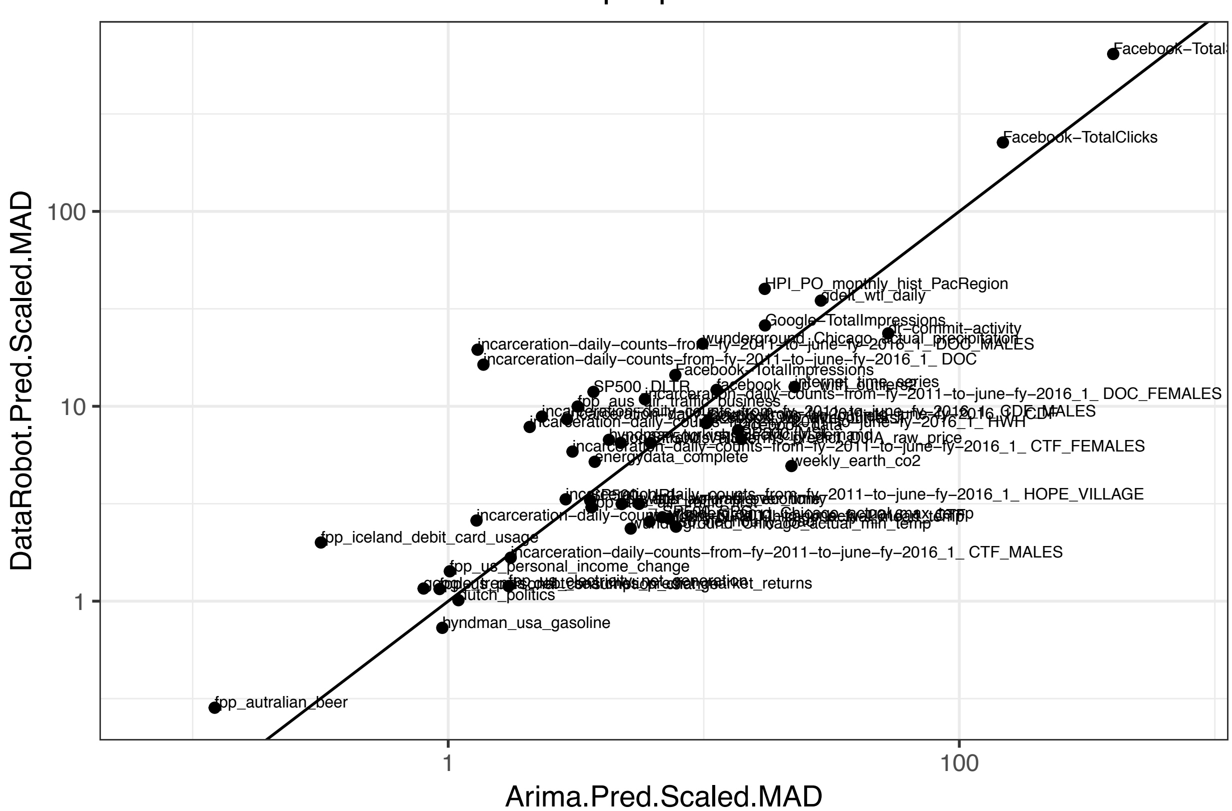


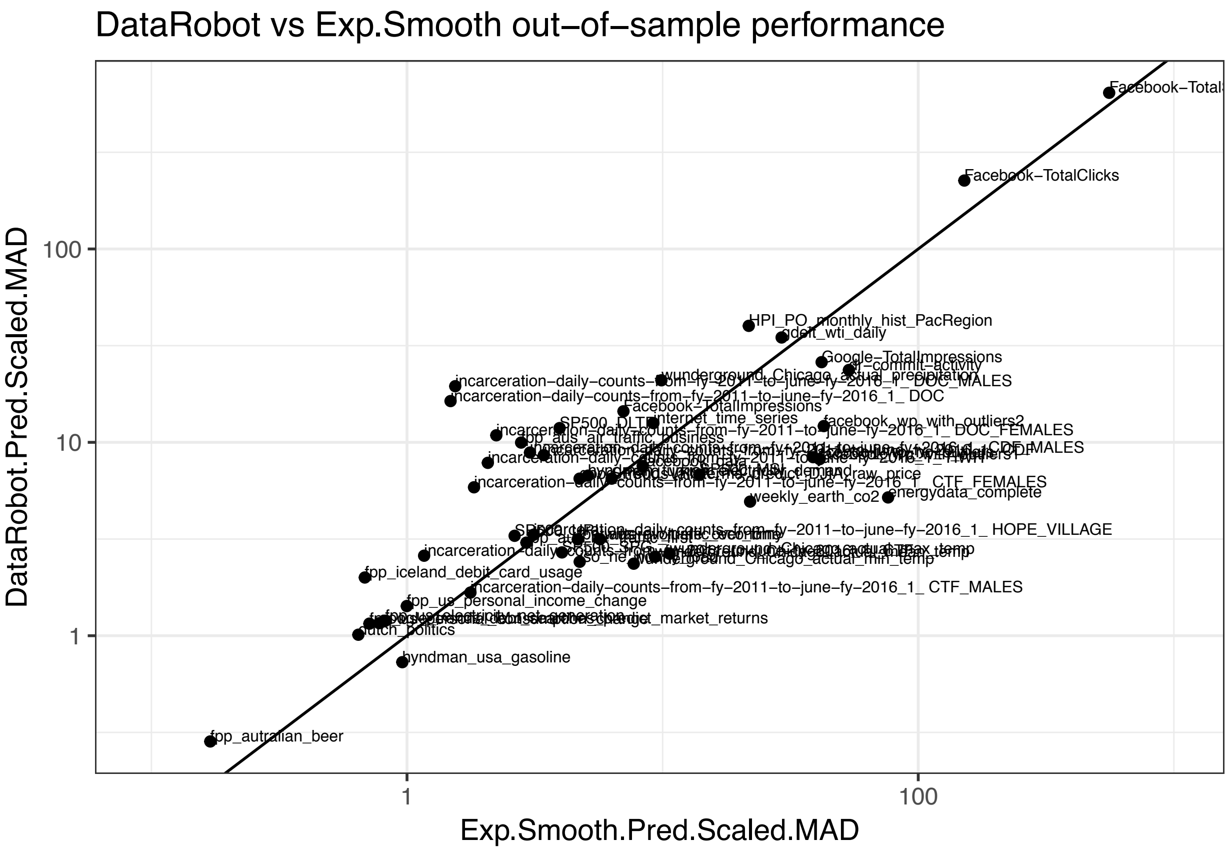
DataRobot vs Prophet out−of−sample performance
DataRobot.Pred.Scaled.MAD
Prophet.Pred.Scaled.MAD
1
10
100
Facebook−TotalClicks
gdelt_wti_daily
Google−TotalImpressions
dr−commit−activity
wunderground_Chicago_actual_precipitation
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ DOC
Facebook−TotalImpressions
internet_time_series
facebook_wp_with_outliers2
fpp_aus_air_traffic_business
DJIA_raw_price
energydata_complete
weekly_earth_co2
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ HOPE_VILLAGE
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ CTF
wunderground_Chicago_actual_min_temp
fpp_iceland_debit_card_usage
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ CTF_MALES
fpp_us_personal_income_change
fpp_us_electricity_net_generation
dutch_politics
hyndman_usa_gasoline
fpp_autralian_beer

DataRobot vs TBATS out–of–sample performance



DataRobot vs Arima out–of–sample performance




DataRobot vs Exp.Smooth out−of−sample performance
Facebook−TotalClicks
HPI_PO_monthly_hist_PacRegion
Google−TotalImpressions
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ DOC
Facebook−TotalImpressions
internet_time_series
facebook_wp_with_outliers2
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ DOC_FEMALES
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ CTF_FEMALES
weekly_earth_co2
energydata_complete
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ HOPE_VILLAGE
wunderground_Chicago_actual_min_temp
fpp_iceland_debit_card_usage
incarceration−daily−counts−from−fy−2011−to−june−fy−2016_1_ CTF_MALES
fpp_us_personal_income_change
dutch_politics
hyndman_usa_gasoline
fpp_autralian_beer
100
10
1
1
100
DataRobot.Pred.Scaled.MAD
Exp.Smooth.Pred.Scaled.MAD

	dataset	DataRobot.Pred.Scaled.MAD	Eureqa.Pred.Scaled.MAD	Prophet.Pred.Scaled.MAD	TBATS.Pred.Scaled.MAD	Arima.Pred.Scaled.MAD	Exp.Smooth.Pred.Scaled.MAD	DR.Min.Improve
1	Facebook-TotalClicks	225.8	1185.6	174.5	160.1	148.1	151.5	959.8
2	Facebook-TotalSpend	642.0	1313.1	242.6	464.0	400.2	558.4	671.0
3	internet_time_series	12.6	9.3	12.4	356.7	22.7	9.2	344.2
4	weekly_earth_co2	4.9	188.5	4.0	17.7	22.0	22.0	183.6
5	SP500_MSI	6.8	135.0	4.5	13.2	14.0	13.9	128.2
6	HPI_PO_monthly_hist_PacRegion	40.0	118.2	140.2	10.0	17.3	21.7	100.1
7	gdelt_wti_daily	34.8	134.2	9.5	33.0	28.7	29.2	99.4
8	SP500_VRSK	6.5	84.3	7.2	4.0	6.2	6.3	77.8
9	energydata_complete	5.2	3.8	2.9	3.2	3.7	76.2	71.0
10	SP500_URI	3.3	74.1	2.2	3.0	3.6	2.6	70.8
11	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ CDF	8.6	66.5	18.4	10.7	2.9	3.4	57.9
12	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ CDF_MALES	8.9	65.7	19.0	12.3	2.3	3.0	56.9
13	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ HWH	7.8	50.3	8.3	2.3	2.1	2.1	42.5
14	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ DOC_FEMALES	10.9	48.4	15.5	1.9	5.9	2.2	37.5
15	facebook_data	7.5	41.5	16.5	8.3	13.6	8.4	33.9
16	facebook_wp_with_outliers1	8.2	31.8	9.5	10.7	10.2	41.1	32.9
17	facebook_wp_with_outliers2	12.1	32.9	10.7	11.1	11.2	42.7	30.6
18	facebook_wp_no_outliers	8.4	34.8	9.0	10.5	10.4	38.8	30.4
19	dr-commit-activity	23.7	18.6	21.9	34.4	52.7	53.6	29.9
20	Facebook-TotalImpressions	14.5	41.4	5.3	9.7	7.7	7.0	27.0
21	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ DOC	16.3	41.7	10.1	4.2	1.4	1.5	25.3
22	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ CTF_FEMALES	5.9	29.9	8.8	1.5	3.1	1.8	24.1
23	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ DOC_MALES	19.5	43.2	10.8	1.6	1.3	1.5	23.7
24	Google-TotalImpressions	26.0	46.6	19.9	27.9	17.4	41.9	20.7
25	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ CTF	2.6	16.9	4.4	1.5	1.3	1.2	14.3
26	hyndman_turkish_electricity_demand	6.7	19.5	2.4	4.6	4.3	5.1	12.8
27	SP500_SPG	2.7	12.1	3.3	5.4	6.9	4.0	9.4
28	hyndman_usa_gasoline	0.7	10.0	0.5	0.7	0.9	1.0	9.3
29	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ CTF_MALES	1.7	10.3	3.2	2.0	1.8	1.8	8.6
30	wunderground_Chicago_actual_max_temp	2.7	2.5	2.3	7.3	7.3	10.7	8.0
31	wunderground_Chicago_actual_mean_temp	2.6	2.1	1.9	6.1	6.1	9.3	6.8
32	twitter_volume_over_time	3.2	3.7	2.9	9.7	5.6	5.7	6.6
33	iso_ne_hourly_load	2.4	2.5	3.0	6.5	7.8	4.7	5.4
34	wunderground_Chicago_actual_min_temp	2.4	1.8	1.8	5.2	5.2	7.7	5.3
35	fpp_aus_air_traffic_economy	3.2	7.7	2.6	2.4	4.8	4.7	4.6
36	fpp_aus_air_traffic_first	3.0	3.5	5.6	4.9	3.6	2.9	2.6
37	fpp_aus_air_traffic_business	10.0	10.0	12.3	2.8	3.2	2.8	2.4
38	dutch_politics	1.0	2.4	0.9	1.8	1.1	0.6	1.4
39	fpp_us_electricity_net_generation	1.2	2.1	1.8	0.8	1.7	0.8	0.9
40	fpp_autralian_beer	0.3	0.2	0.2	0.3	0.1	0.2	0.0
41	goog_trends_all_terms_predict_DJIA_raw_price	6.5	6.4	3.1	4.7	4.7	4.7	-0.1
42	fpp_iceland_debit_card_usage	2.0	1.8	1.4	1.3	0.3	0.7	-0.2
43	incarceration-daily-counts-from-fy-2011-to-june-fy-2016_1_ HOPE_VILLAGE	3.3	2.7	2.9	2.7	2.9	3.1	-0.2
44	fpp_us_personal_consumption_change	1.2	0.8	0.9	0.7	0.9	0.7	-0.2
45	google_trends_debt_searches_predict_market_returns	1.2	0.8	0.9	0.8	0.8	0.8	-0.3
46	fpp_us_personal_income_change	1.4	1.0	1.0	1.0	1.0	1.0	-0.4
47	SP500_DLTR	11.9	7.0	10.9	3.4	3.7	4.0	-1.0
48	wunderground_Chicago_actual_precipitation	20.9	10.3	14.1	10.8	9.9	9.9	-6.8

Table 2: Full Results