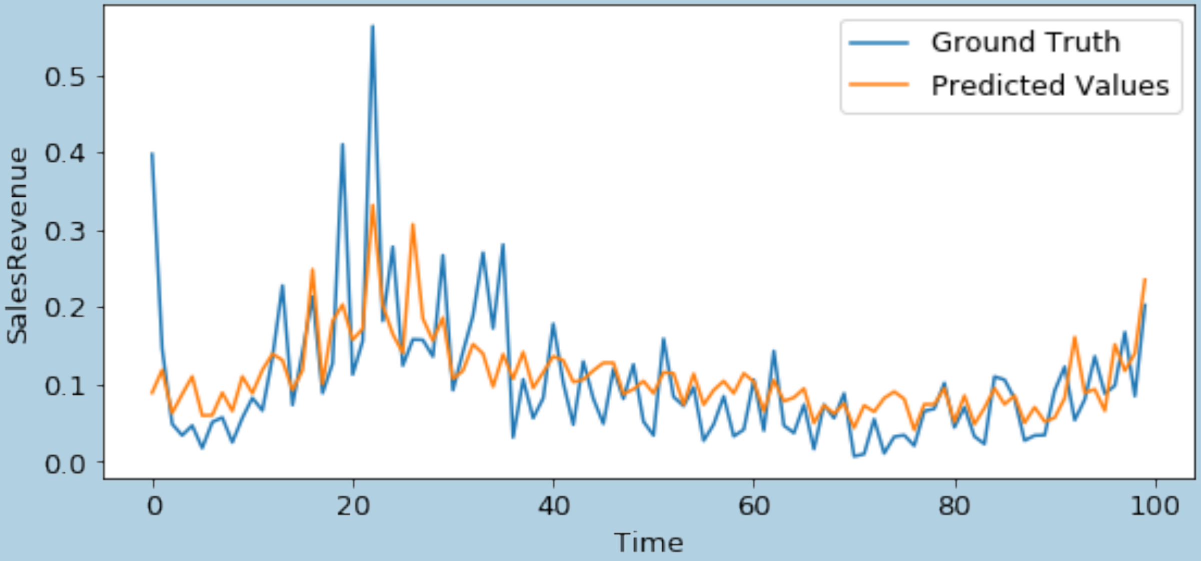
SALES FORECASTING WITH MACHINE LEARNING

ALEX REIBMAN



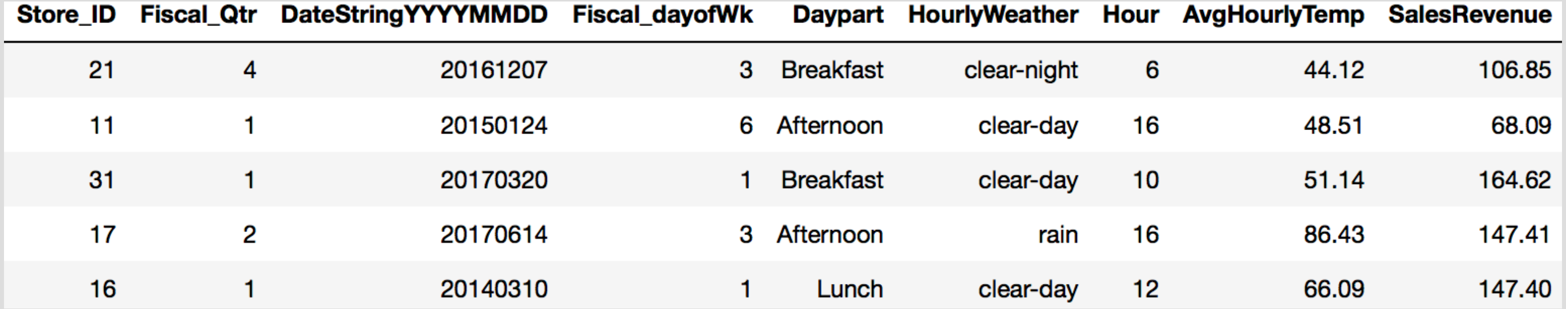
OBJECTIVE

- Analyze time series sales data for 14 quick-serve food chains
- Identify trends in purchasing over time
- Determine key periods of performance
- Identify top predictors
- Create forecast for future sales

DATA MODELING PROCESS

- One-hot encode categorical features
- Remove negative and large outliers from training set
- Include rolling averages for the entire franchise as well as each individual store
- Scale data in values between 0 and 1
- Test 5 separate models:
  - General goodness of fit
  - Rolling window cross validation
- Two separate modeling processes:
  - Model all data available across all stores and make predictions
  - Model data unique to each store and make predictions

DATA DESCRIPTION

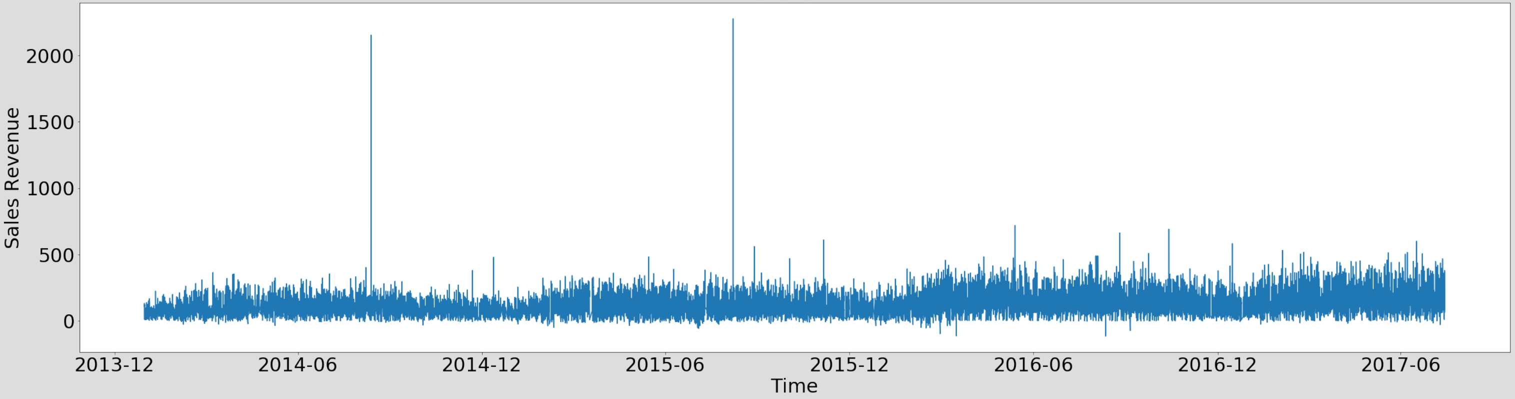


Store_ID	Fiscal_Qtr	DateStringYYYYMMDD	Fiscal_dayofWk	Daypart	HourlyWeather	Hour	AvgHourlyTemp	SalesRevenue
21	4	20161207	3	Breakfast	clear-night	6	44.12	106.85
11	1	20150124	6	Afternoon	clear-day	16	48.51	68.09
31	1	20170320	1	Breakfast	clear-day	10	51.14	164.62
17	2	20170614	3	Afternoon	rain	16	86.43	147.41
16	1	20140310	1	Lunch	clear-day	12	66.09	147.40

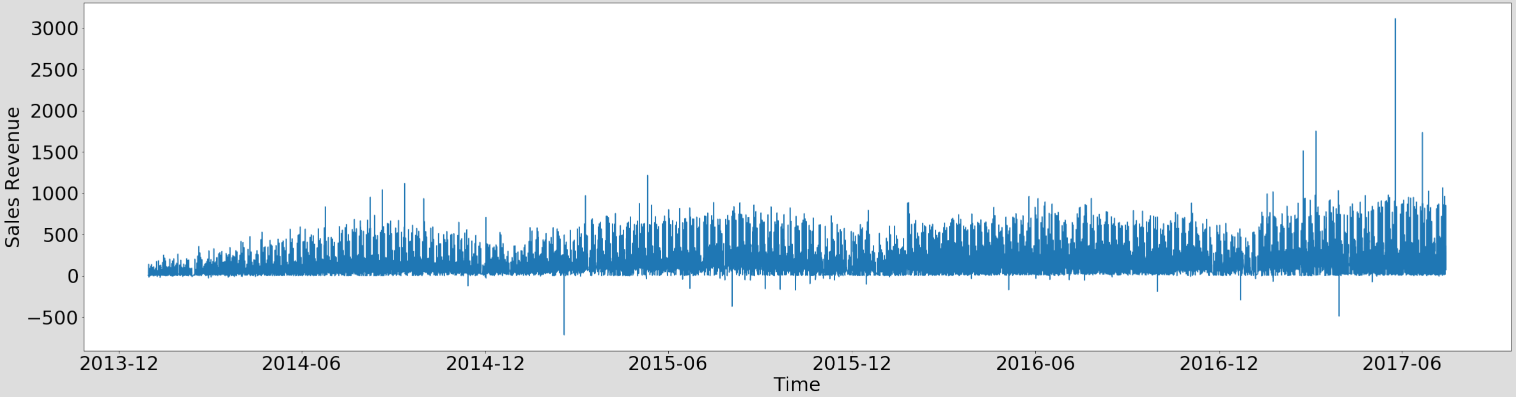
- Store ID- Unique ID of store
- Fiscal Quarter
- Date and day of week
- Part of day
- Weather
- Temperature
- Sales Revenue

SAMPLES

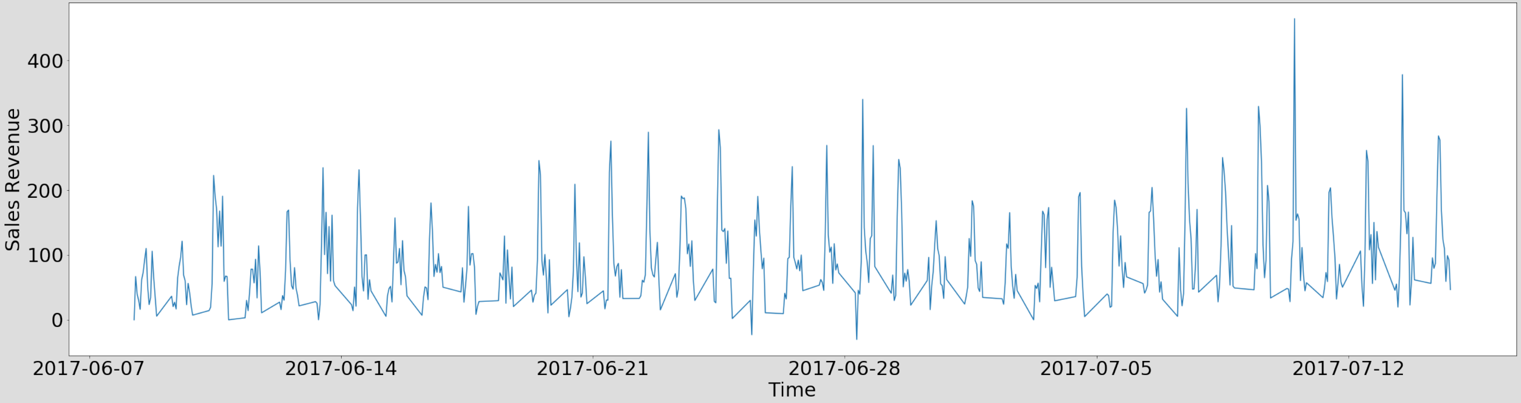
Store 2



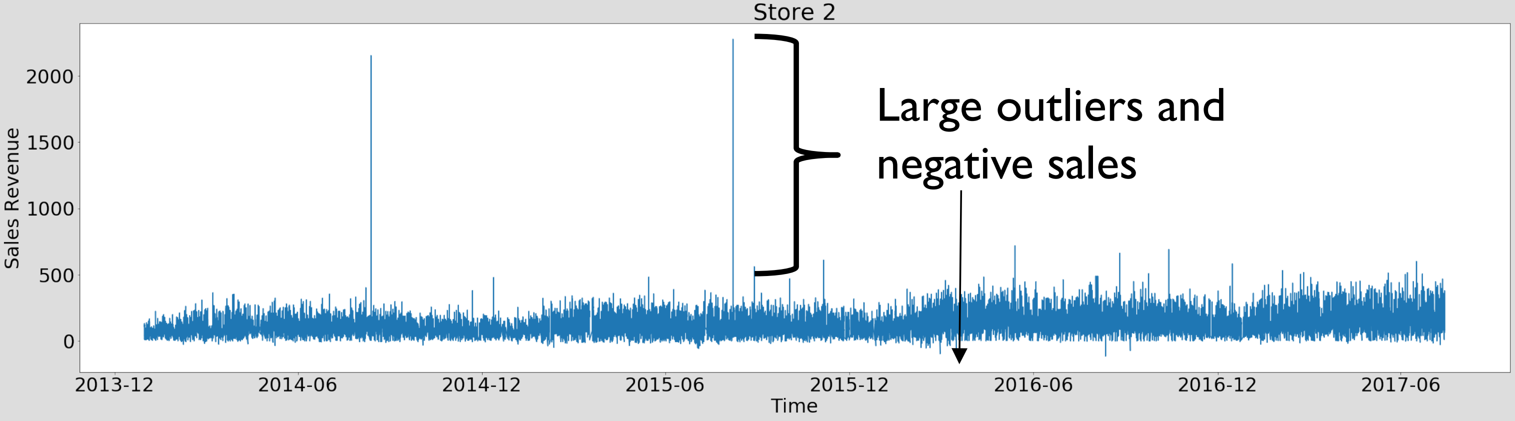
Store 17



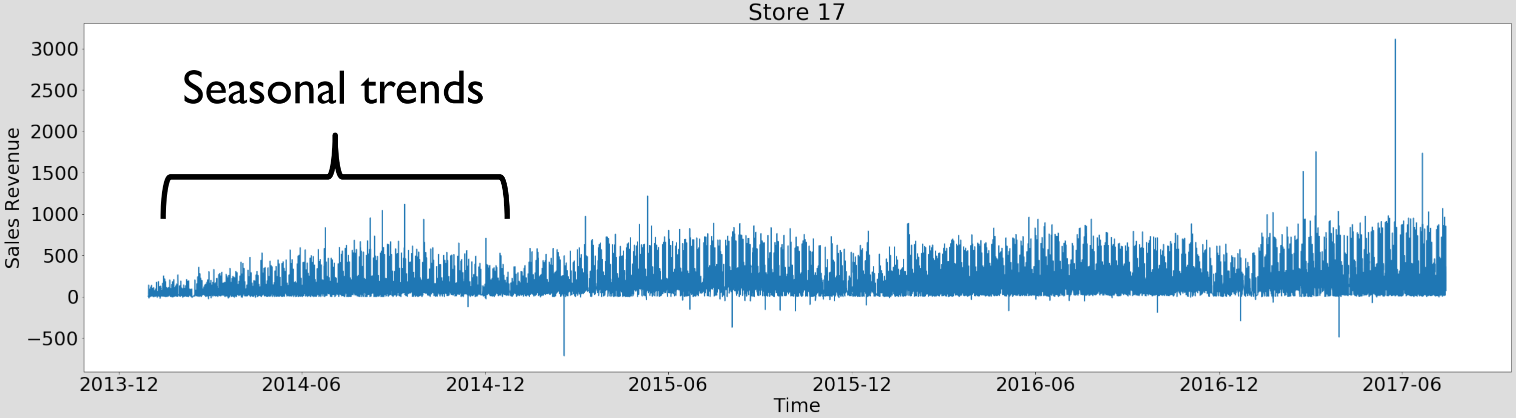
Store 23



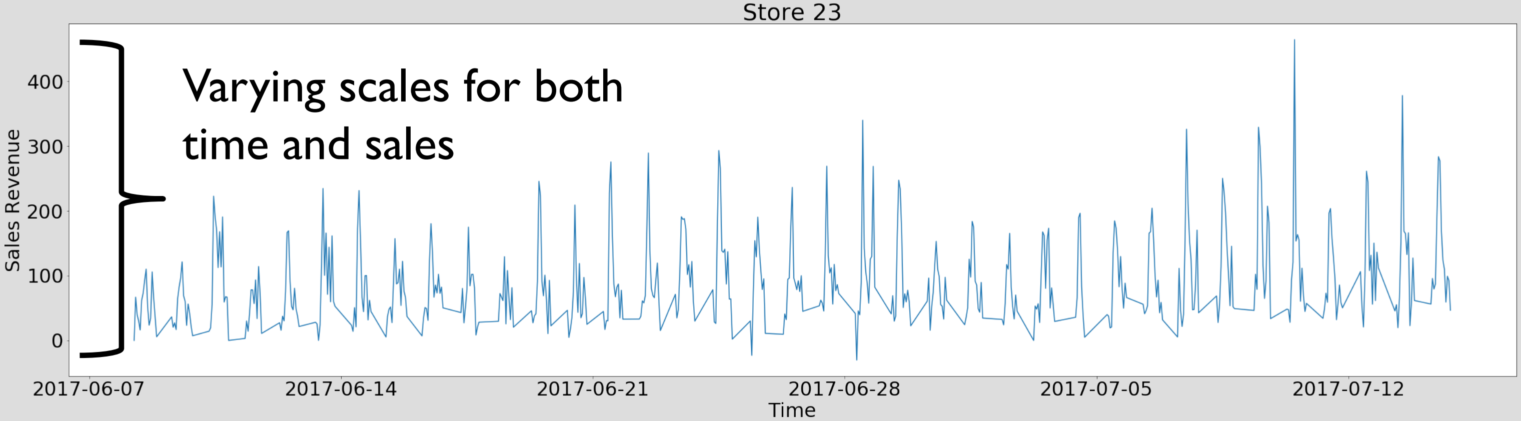
PROBLEMS



Large outliers and negative sales

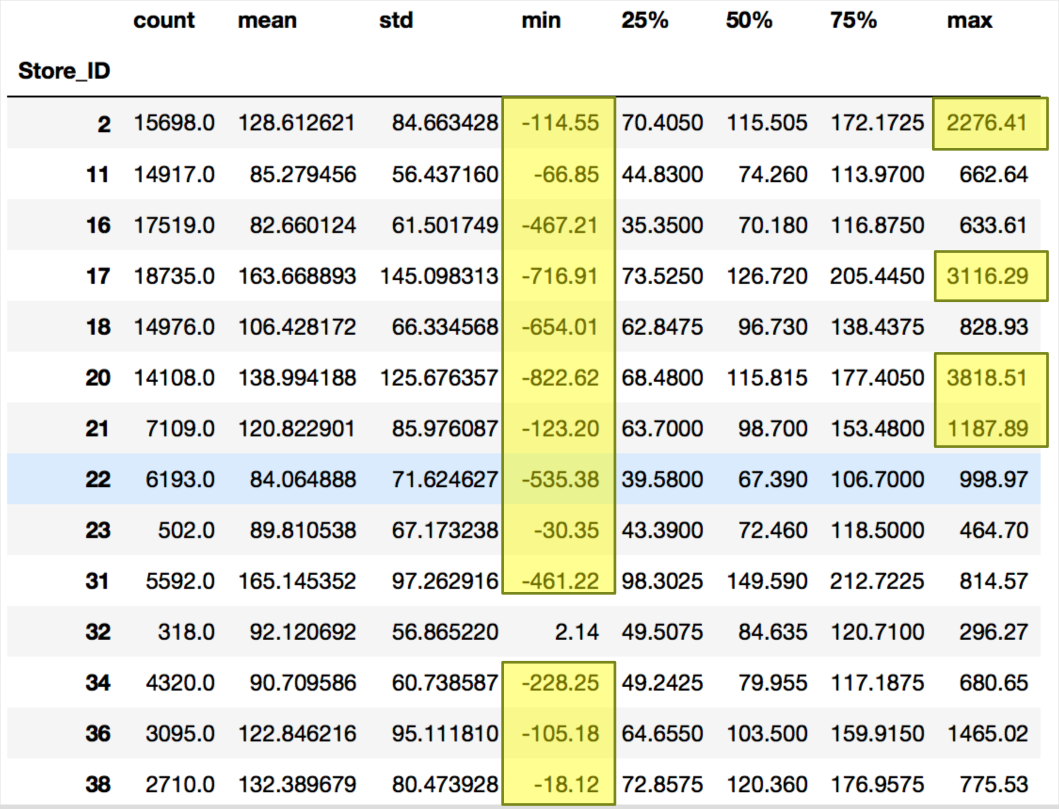


Seasonal trends

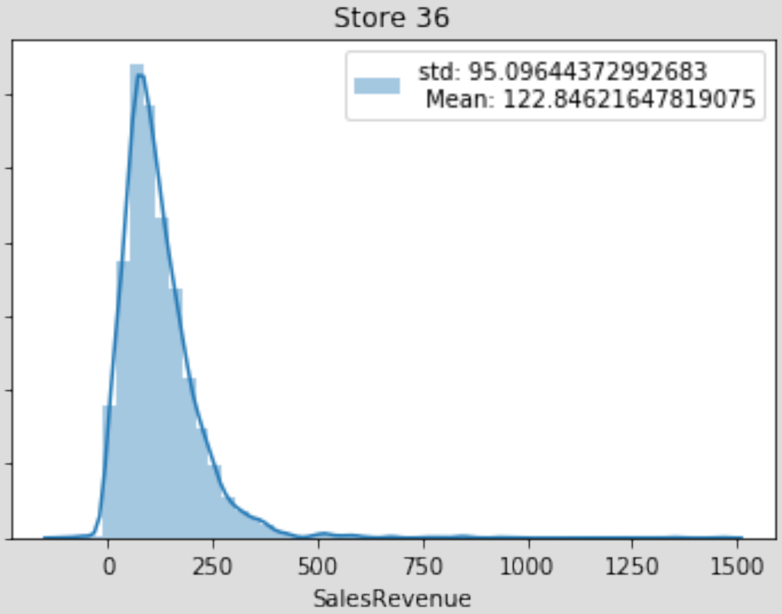


Varying scales for both time and sales

DISTRIBUTIONS



	count	mean	std	min	25%	50%	75%	max
Store_ID								
2	15698.0	128.612621	84.663428	-114.55	70.4050	115.505	172.1725	2276.41
11	14917.0	85.279456	56.437160	-66.85	44.8300	74.260	113.9700	662.64
16	17519.0	82.660124	61.501749	-467.21	35.3500	70.180	116.8750	633.61
17	18735.0	163.668893	145.098313	-716.91	73.5250	126.720	205.4450	3116.29
18	14976.0	106.428172	66.334568	-654.01	62.8475	96.730	138.4375	828.93
20	14108.0	138.994188	125.676357	-822.62	68.4800	115.815	177.4050	3818.51
21	7109.0	120.822901	85.976087	-123.20	63.7000	98.700	153.4800	1187.89
22	6193.0	84.064888	71.624627	-535.38	39.5800	67.390	106.7000	998.97
23	502.0	89.810538	67.173238	-30.35	43.3900	72.460	118.5000	464.70
31	5592.0	165.145352	97.262916	-461.22	98.3025	149.590	212.7225	814.57
32	318.0	92.120692	56.865220	2.14	49.5075	84.635	120.7100	296.27
34	4320.0	90.709586	60.738587	-228.25	49.2425	79.955	117.1875	680.65
36	3095.0	122.846216	95.111810	-105.18	64.6550	103.500	159.9150	1465.02
38	2710.0	132.389679	80.473928	-18.12	72.8575	120.360	176.9575	775.53



- Large irregularities lead to skewed distributions.
- Negative values present

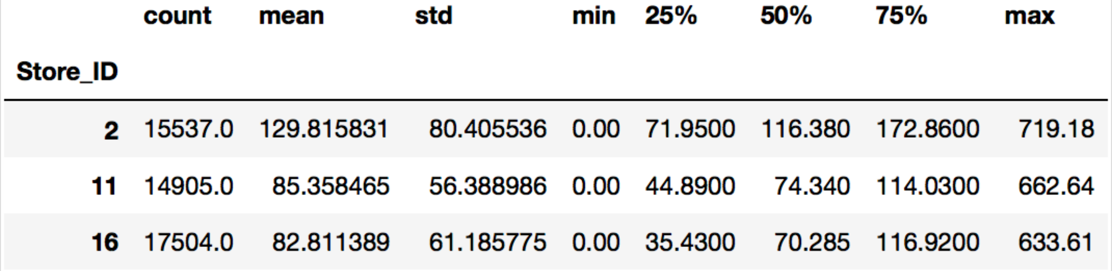
TRANSFORM CATEGORICAL VARIABLES INTO DUMMY VARIABLES

The following features were one-hot encoded:

- Fiscal_Qtr- A dummy variable for each quarter
- Fiscal_dayofWk- 1 through 7
- Daypart- Breakfast, lunch, dinner, etc.
- HourlyWeather- Clear-day, rain, etc.
- Store_ID

Holidays did not seem to vary enough to warrant encoding.

MODIFICATIONS

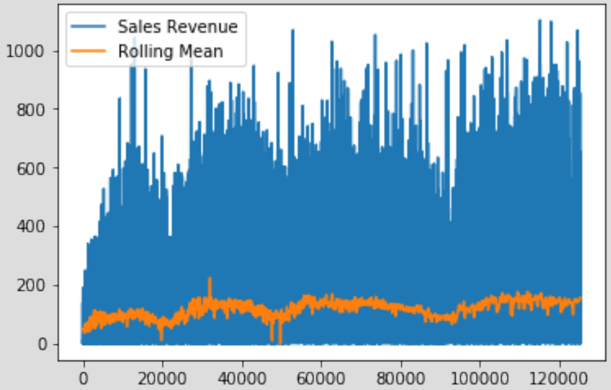


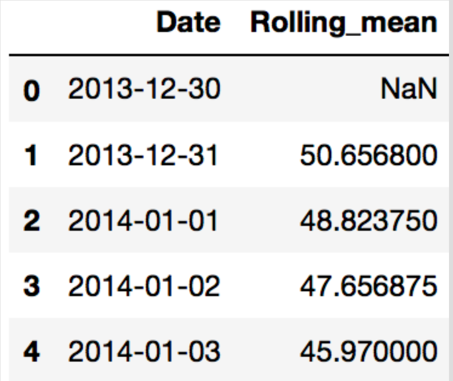
	count	mean	std	min	25%	50%	75%	max
Store_ID								
2	15537.0	129.815831	80.405536	0.00	71.9500	116.380	172.8600	719.18
11	14905.0	85.358465	56.388986	0.00	44.8900	74.340	114.0300	662.64
16	17504.0	82.811389	61.185775	0.00	35.4300	70.285	116.9200	633.61

- Negative values removed
- Any value greater 10 standard deviations from the mean removed

ROLLING MEAN

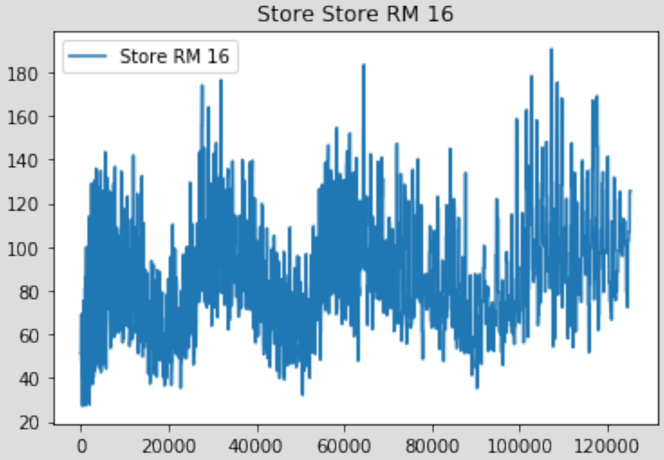
- Encode average daily sales of all stores from previous day





	Date	Rolling_mean
0	2013-12-30	NaN
1	2013-12-31	50.656800
2	2014-01-01	48.823750
3	2014-01-02	47.656875
4	2014-01-03	45.970000

- Average daily sales from each individual store



PREDICTIVE MODELS

Categories of models

- **Generalized models-** fit on sales revenue for all stores
- **Store specific models-** fit on sales revenue for each individual store

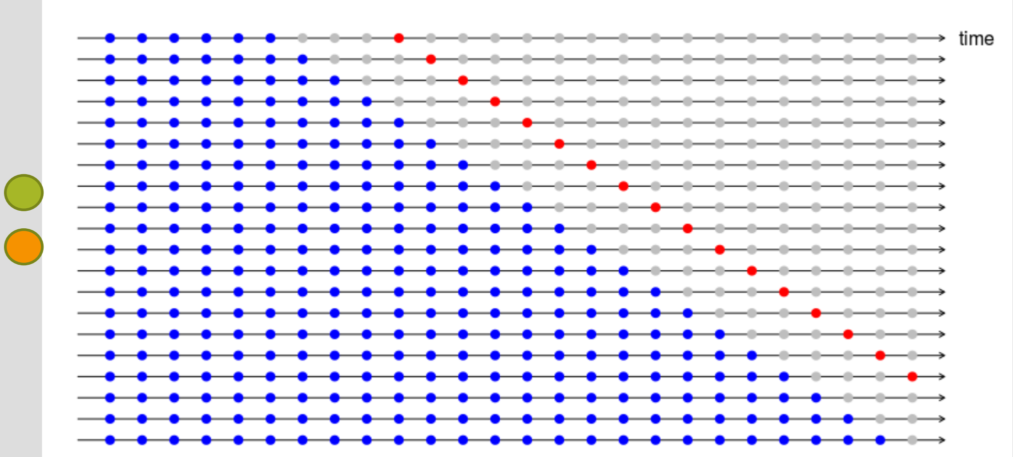
PREDICTIVE MODELS

- 5 separate models selected:
- **Linear models:**
  - Linear regression (ordinary least squares)
  - Ridge Regression
  - Lasso Regression
- **Decision tree based models-**
  - Random Forest Regression
  - Gradient Boosting Regression
- Models were chosen on a basis of interpretability, intuitiveness, and performance.

METRICS:

Models were tested by two categories:

- General goodness of fit in R^2
  - How well the model fits to all recorded data
- Rolling window cross validation
  - Each dataset was split into 10 separate time windows. The model is fitted on 10% of the earliest samples and tested on the remaining samples. Then, fitted on 20%, etc.

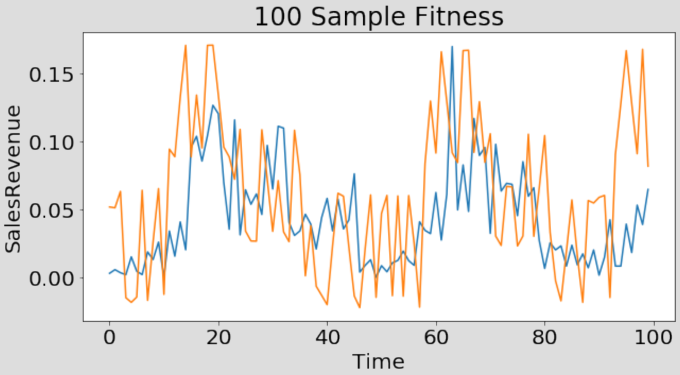


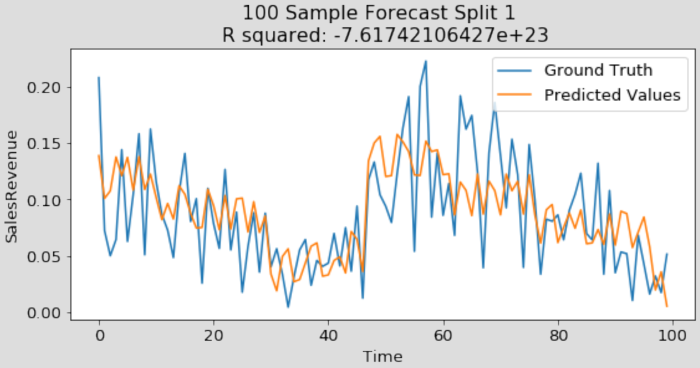
Training set 🟢

Testing set 🟠

GENERALIZED MODELS

LINEAR REGRESSION RESULTS: POOR


100 Sample Fitness

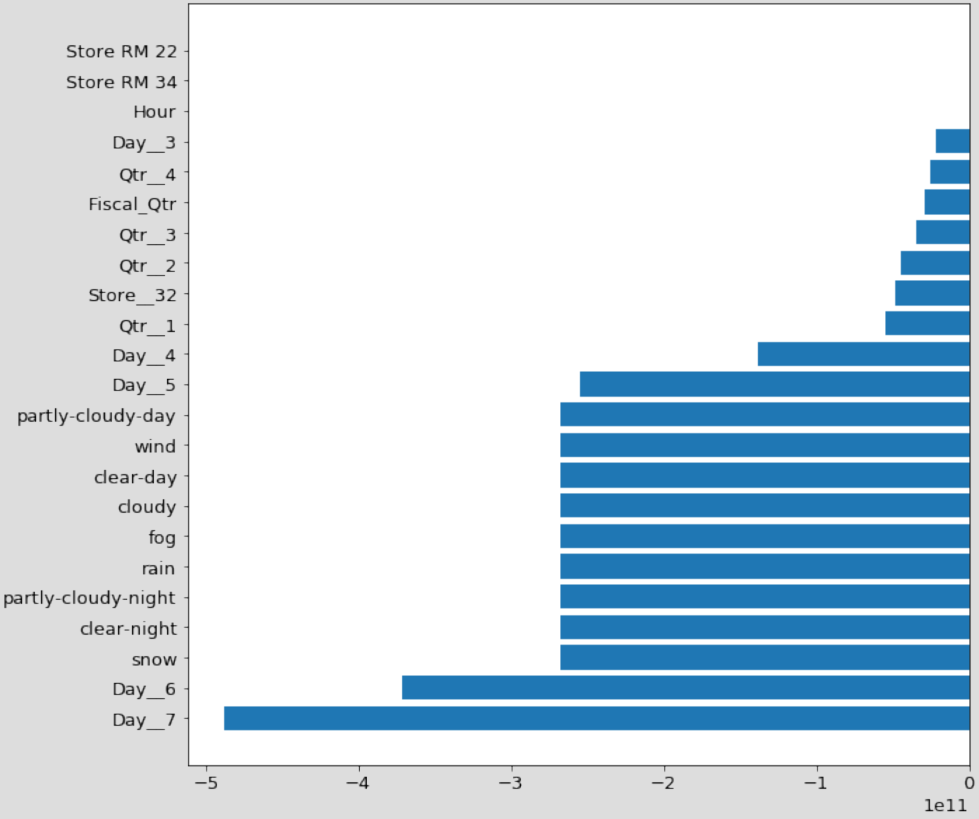

100 Sample Forecast Split 1
R squared: -7.61742106427e+23

- Split 1 R squared: -7.61742106427e+23
- Split 2 R squared: 0.3593
- Split 3 R squared: 0.3555
- Split 4 R squared: -3.69375046051e+21
- Split 5 R squared: -3.7446634236e+21
- Split 6 R squared: -3.81862786533e+22
- Split 7 R squared: -4.05582421931e+21
- Split 8 R squared: -2.85868947422e+21
- Split 9 R squared: 0.432
- Split 10 R squared: -3.85342116196e+24

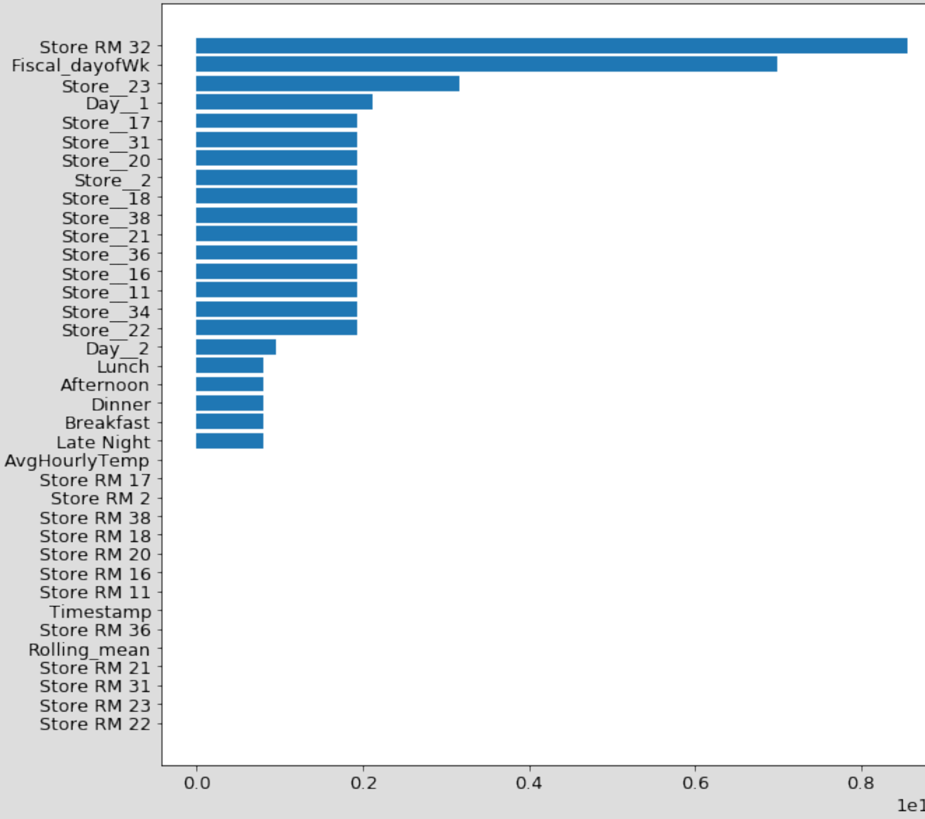
- R squared 0.45839

COEFFICIENTS

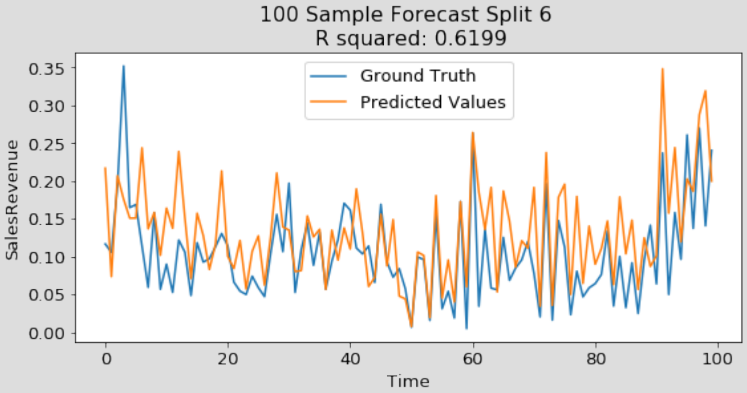
Negative Values

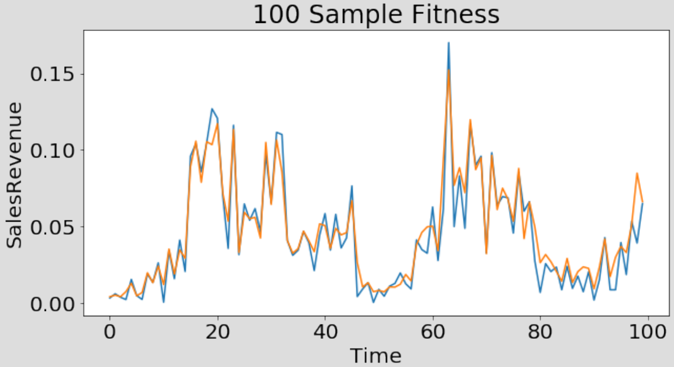


Positive Values



RANDOM FOREST REGRESSION RESULTS: GOOD

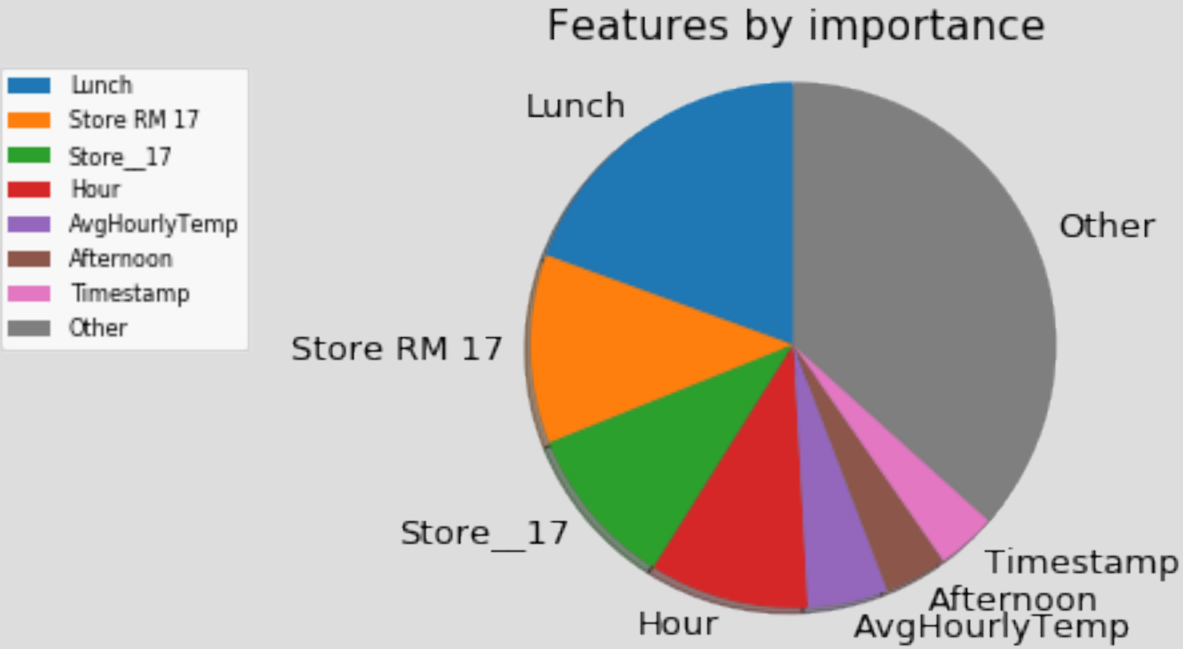




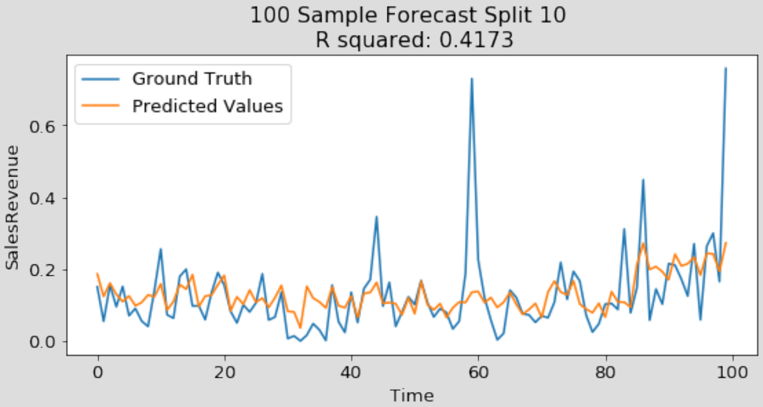
R squared 0.9413

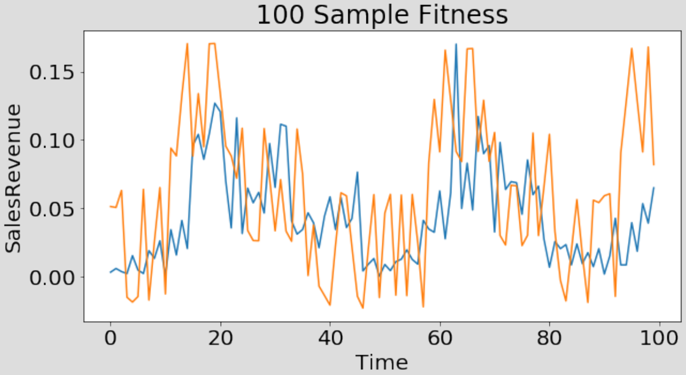
- Split 1 R squared: 0.535
- Split 2 R squared: 0.4293
- Split 3 R squared: 0.6417
- Split 4 R squared: 0.5489
- Split 5 R squared: 0.5941
- Split 6 R squared: 0.6199
- Split 7 R squared: 0.5894
- Split 8 R squared: 0.5616
- Split 9 R squared: 0.5916
- Split 10 R squared: 0.618

FEATURES BY IMPORTANCE



RIDGE REGRESSION RESULTS: DECENT

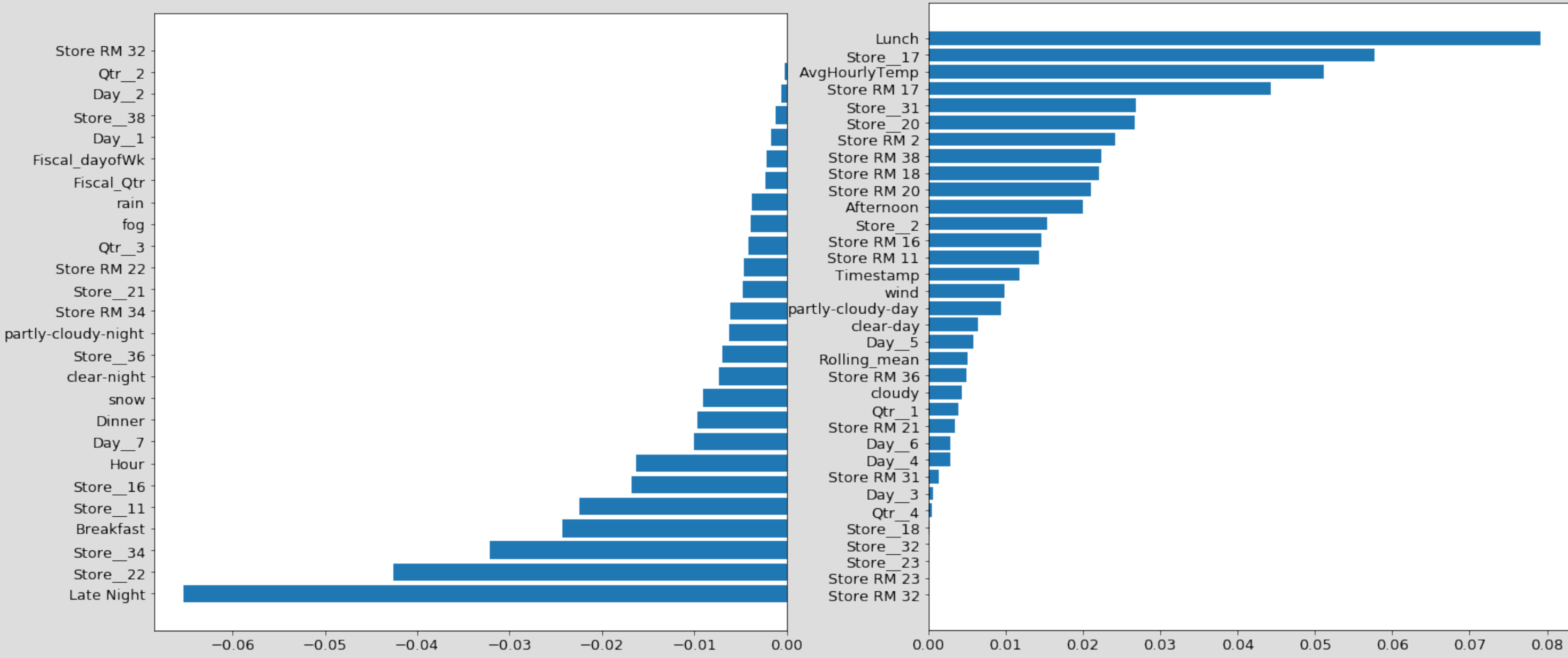




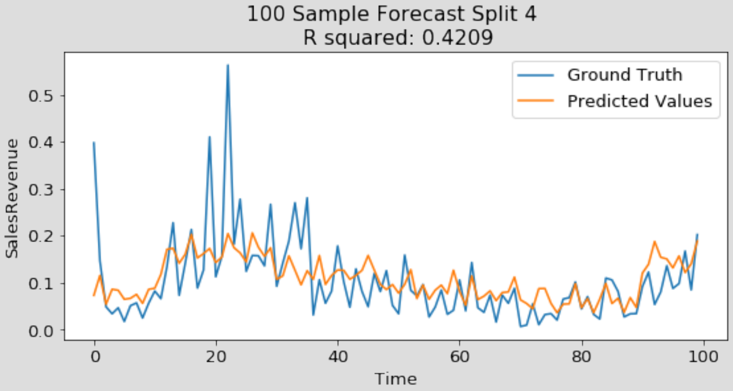
- Split 1 R squared: 0.2146
- Split 2 R squared: 0.3596
- Split 3 R squared: 0.3551
- Split 4 R squared: 0.4226
- Split 5 R squared: 0.3997
- Split 6 R squared: 0.4531
- Split 7 R squared: 0.4273
- Split 8 R squared: 0.4547
- Split 9 R squared: 0.4316
- Split 10 R squared: 0.4173

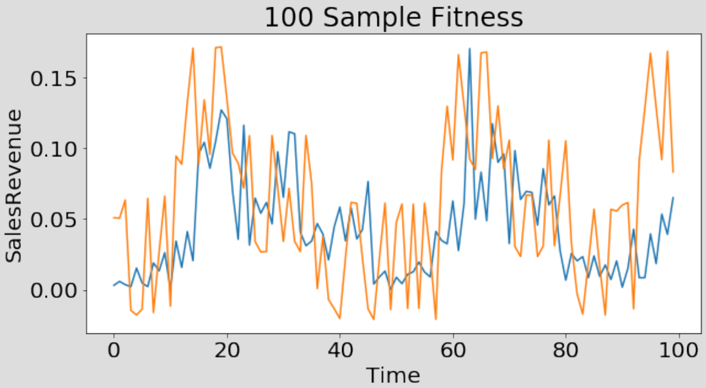
- R squared 0.4584

COEFFICIENTS



LASSO REGRESSION RESULTS: DECENT

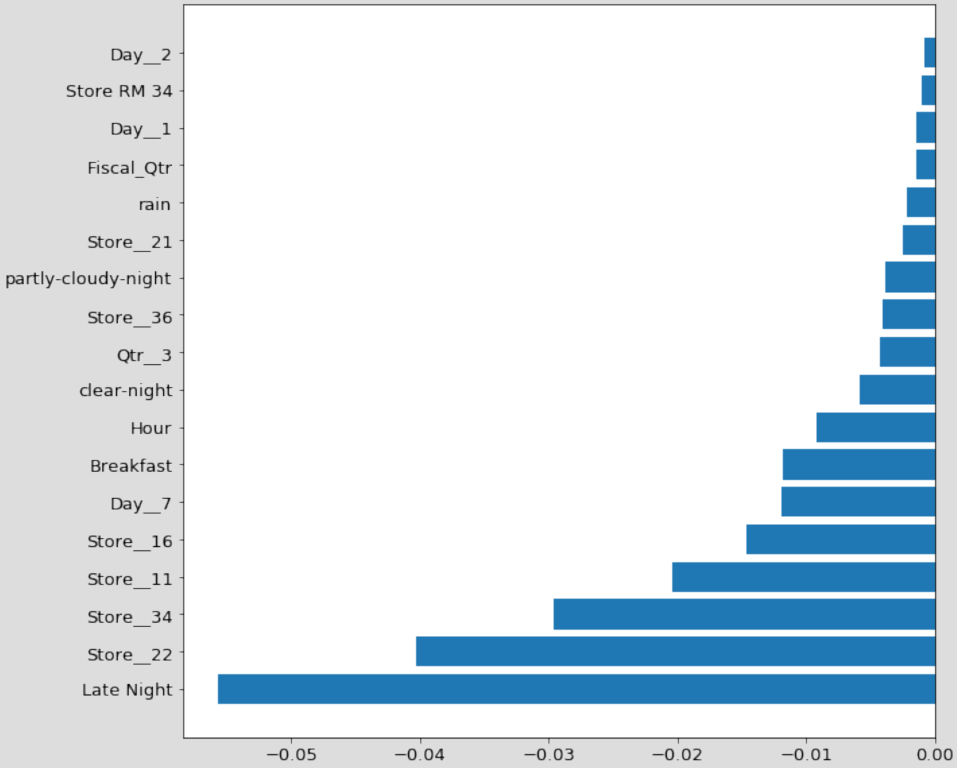


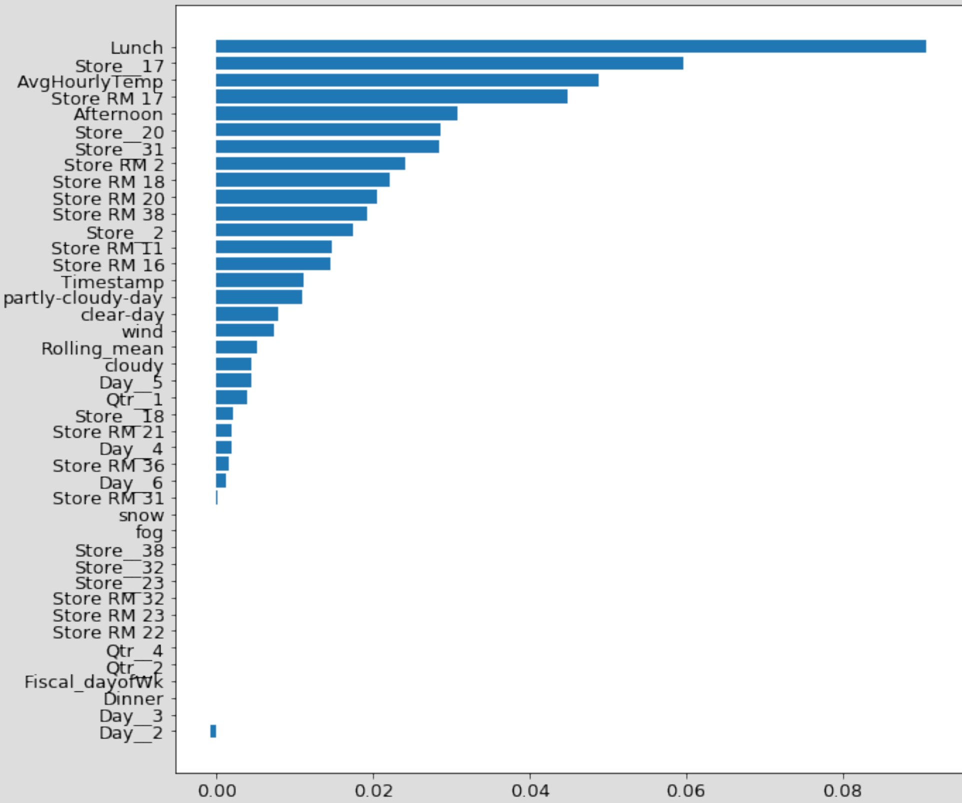


- Split 1 R squared: 0.2739
- Split 2 R squared: 0.2876
- Split 3 R squared: 0.3554
- Split 4 R squared: 0.4209
- Split 5 R squared: 0.4083
- Split 6 R squared: 0.454
- Split 7 R squared: 0.4342
- Split 8 R squared: 0.4568
- Split 9 R squared: 0.4284
- Split 10 R squared: 0.4186

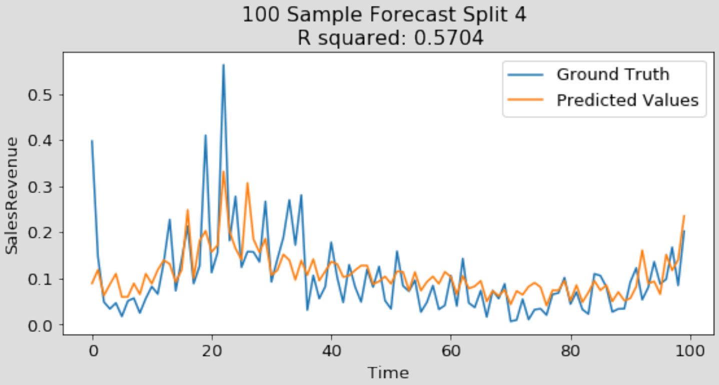
- R squared 0.458

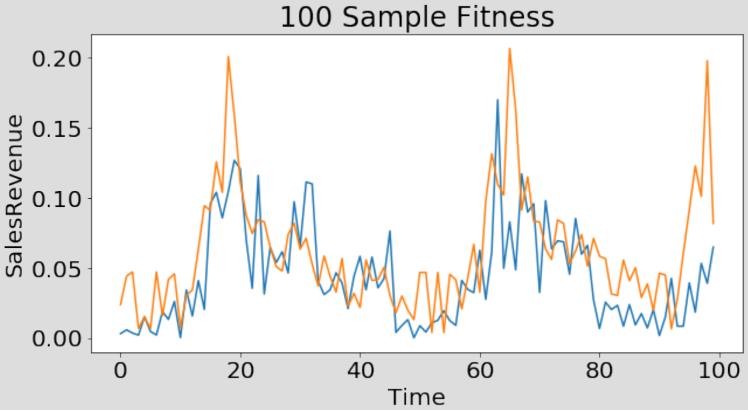
COEFFICIENTS





GRADIENT BOOSTING REGRESSION RESULTS: GOOD

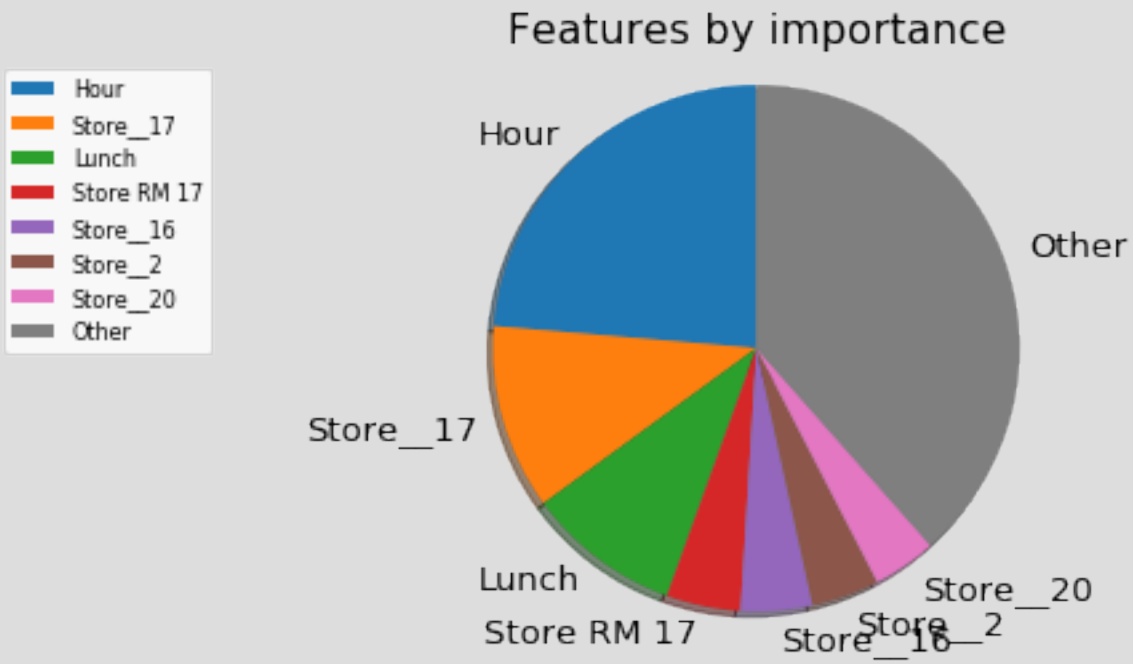




- Split 1 R squared: 0.3999
- Split 2 R squared: 0.4194
- Split 3 R squared: 0.595
- Split 4 R squared: 0.5704
- Split 5 R squared: 0.5841
- Split 6 R squared: 0.6075
- Split 7 R squared: 0.5867
- Split 8 R squared: 0.5935
- Split 9 R squared: 0.5621
- Split 10 R squared: 0.5777

- R squared 0.6318

FEATURES BY IMPORTANCE



SAMPLE PREDICTIONS

- **Store number 17**
- OLS: 170.09197636263153
- Ridge: 170.09197636263153
- Lasso: 92.12069182389936
- Random Forest: 349.986
- GB Regresion: 482.1649243135375

- **Store number 16**
- OLS: 170.09197636263153
- Ridge: 170.09197636263153
- Lasso: 92.12069182389936
- Random Forest: 205.30249999999998
- GB Regresion: 214.63525287681065

STORE SPECIFIC MODELS

INDIVIDUAL STORE PERFORMANCE (STORE 17)

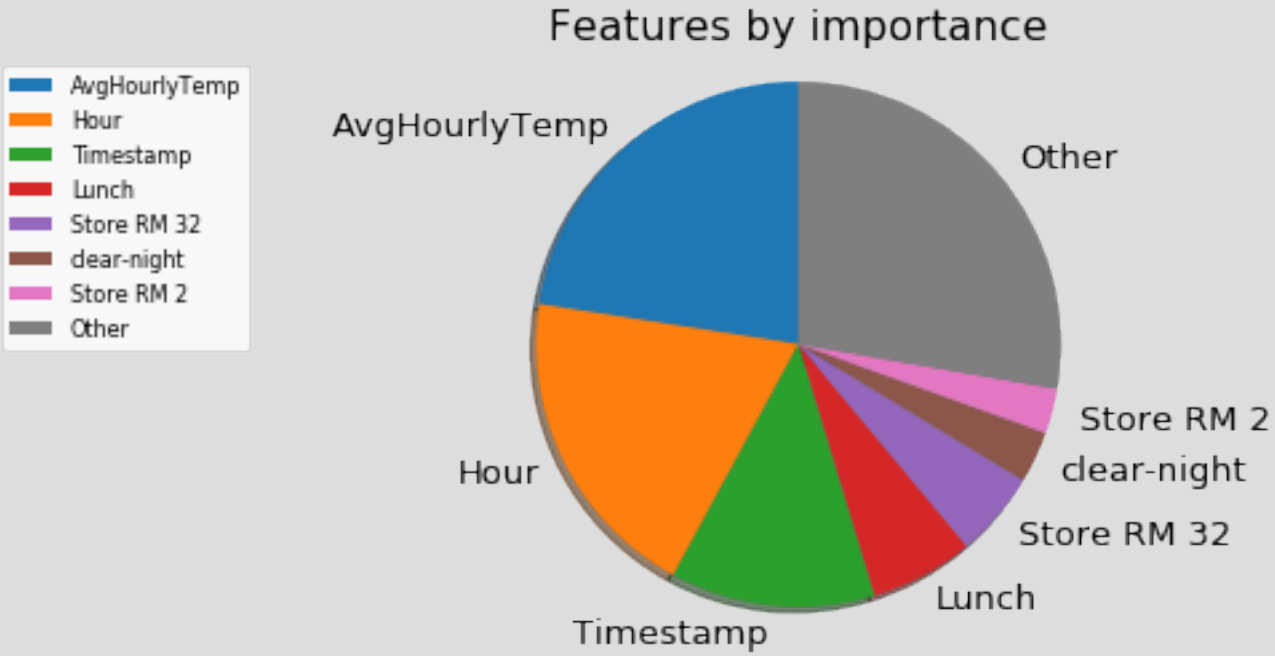
Best Average Performance

Split number	Linear Regression	Lasso	Ridge	Random Forest	Gradient boosting
Split 1	-6.77558824395e+22	0.4263	0.4244	0.6262	0.6326
Split 2	-1.13587982472e+22	0.5175	0.5131	0.6675	0.7608
Split 3	0.5262	0.5121	0.524	0.5708	0.6663
Split 4	0.509	0.5092	0.5083	0.3346	0.7599
Split 5	0.5325	0.5336	0.5322	0.7606	0.7754
Split 6	0.5345	0.5336	0.5344	0.6306	0.6882
Split 7	-4.29203445986e+21	0.5536	0.554	0.7572	0.79
Split 8	-7.95176470223e+22	0.5821	0.5817	0.799	0.8042
Split 9	-3.23627412277e+19	0.551	0.5511	0.6639	0.7632
Split 10	-1.18594223866e+19	0.5134	0.5153	0.7849	0.7833

FEATURES MOST CORRELATED WITH SALES REVENUE

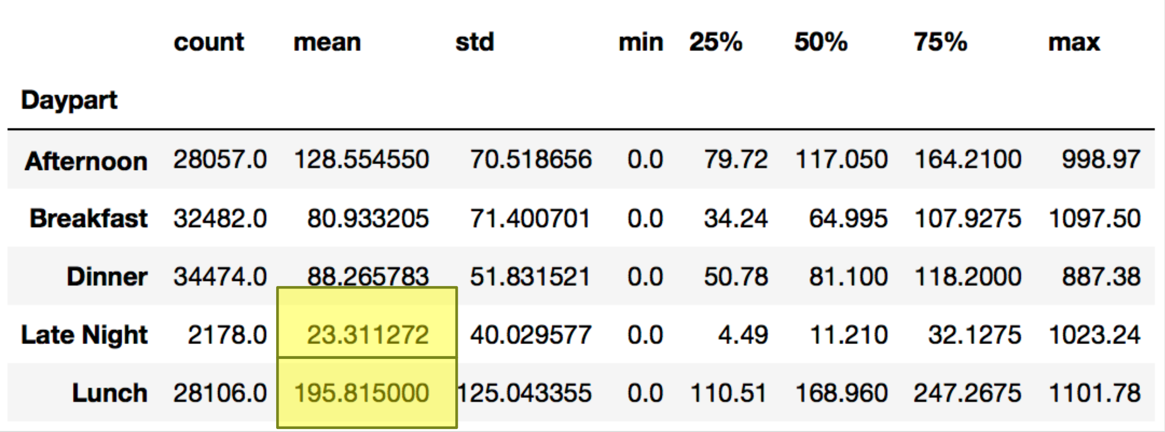
Positive correlation	Negative correlation
Lunch 0.442687	Breakfast -0.235527
Store RM 17 0.251653	Dinner -0.197260
Rolling_mean 0.246261	clear-night -0.195934
Store RM 11 0.239688	Store__16 -0.152389
Store RM 20 0.234312	Late Night -0.134431
Store RM 18 0.230360	Store__11 -0.129006
AvgHourlyTemp 0.229903	Qtr__4 -0.112826
Store RM 2 0.229813	Store__22 -0.082117
Store RM 21 0.208009'	Hour -0.077105
Store__17 0.202817	partly-cloudy-night -0.070913

GRADIENT BOOSTING REGRESSION FEATURE IMPORTANCE



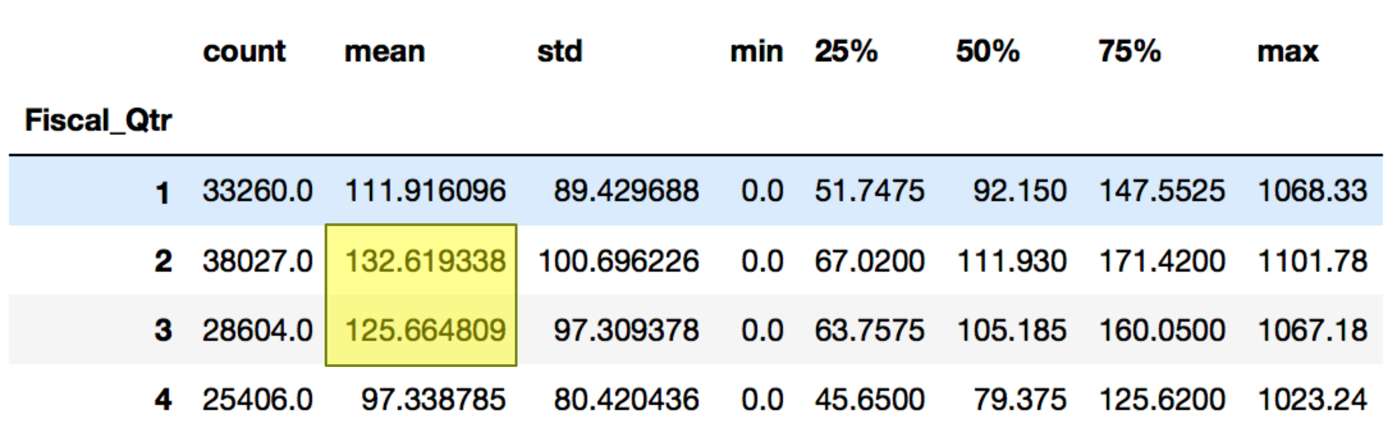
SEASONAL TRENDS

- **Lunchtime leads to higher average sales, yet late night leads to lower sales**



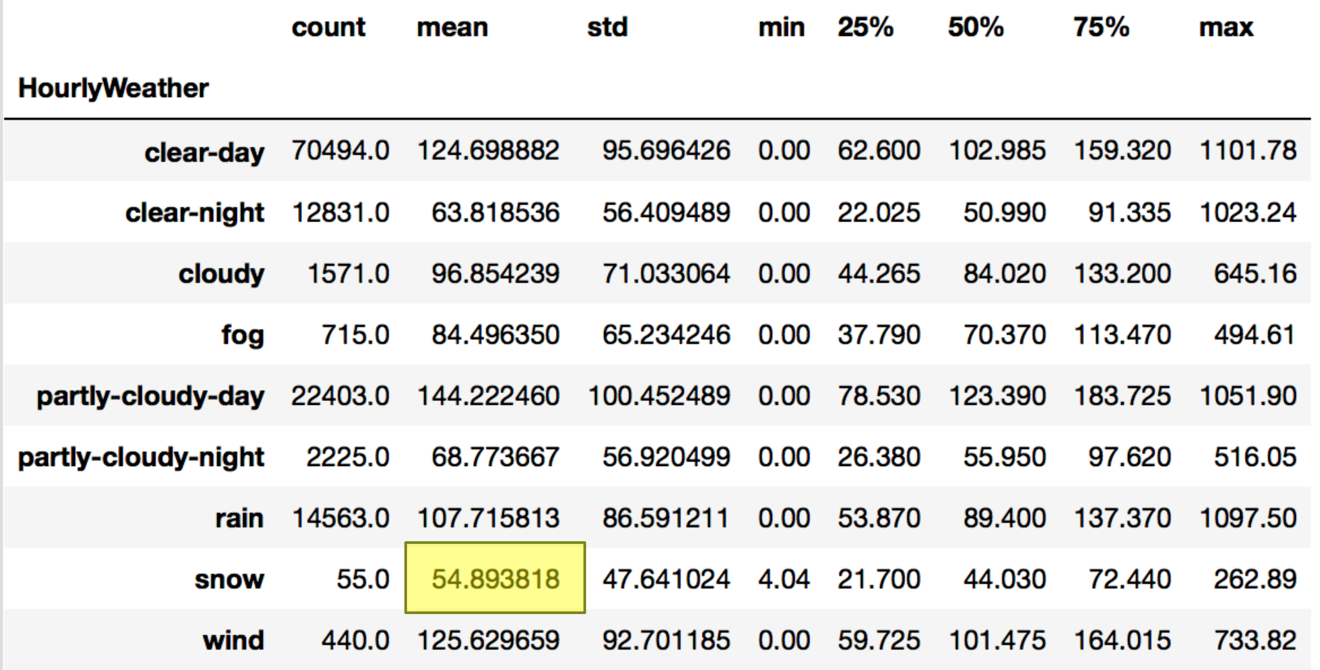
Daypart	count	mean	std	min	25%	50%	75%	max
Afternoon	28057.0	128.554550	70.518656	0.0	79.72	117.050	164.2100	998.97
Breakfast	32482.0	80.933205	71.400701	0.0	34.24	64.995	107.9275	1097.50
Dinner	34474.0	88.265783	51.831521	0.0	50.78	81.100	118.2000	887.38
Late Night	2178.0	23.311272	40.029577	0.0	4.49	11.210	32.1275	1023.24
Lunch	28106.0	195.815000	125.043355	0.0	110.51	168.960	247.2675	1101.78

- **Spring and summer have higher average sales**



Fiscal_Qtr	count	mean	std	min	25%	50%	75%	max
1	33260.0	111.916096	89.429688	0.0	51.7475	92.150	147.5525	1068.33
2	38027.0	132.619338	100.696226	0.0	67.0200	111.930	171.4200	1101.78
3	28604.0	125.664809	97.309378	0.0	63.7575	105.185	160.0500	1067.18
4	25406.0	97.338785	80.420436	0.0	45.6500	79.375	125.6200	1023.24

SEASONAL TRENDS



HourlyWeather	count	mean	std	min	25%	50%	75%	max
clear-day	70494.0	124.698882	95.696426	0.00	62.600	102.985	159.320	1101.78
clear-night	12831.0	63.818536	56.409489	0.00	22.025	50.990	91.335	1023.24
cloudy	1571.0	96.854239	71.033064	0.00	44.265	84.020	133.200	645.16
fog	715.0	84.496350	65.234246	0.00	37.790	70.370	113.470	494.61
partly-cloudy-day	22403.0	144.222460	100.452489	0.00	78.530	123.390	183.725	1051.90
partly-cloudy-night	2225.0	68.773667	56.920499	0.00	26.380	55.950	97.620	516.05
rain	14563.0	107.715813	86.591211	0.00	53.870	89.400	137.370	1097.50
snow	55.0	54.893818	47.641024	4.04	21.700	44.030	72.440	262.89
wind	440.0	125.629659	92.701185	0.00	59.725	101.475	164.015	733.82

SUMMARY

- Gradient boosting regression yields the best performance of the models selected and gives insight as to which features matter the most
- Each store shows a lunch rush trend
- Models fit on data unique to each store perform significantly better
- “Prediction is difficult, especially about the future.” --Niels Bohr