Predicting Retail Electricity Prices

Supervised Machine Learning Approach Using Fossil Fuel Prices from January 2008 to July 2020

Thinkful Capstone Project 2 - Supervised Learning
Michael Dwyer
November 2020

Understanding the Landscape of Power Prices

Industry Structure

Long-term Revenue Contracts and Obligations

The power generation industry operates on a very long-term basis, with most company revenue contracts and fixed asset investment horizons ranging from ten (10) to thirty (30) years long.

Industry Risk

Rapidly Changing Industry with Long-term Contracts

When making large, long-term investments and commitments in an industry with rapidly changing technology and regulation, power generators are exposed to significant risks.

Energy Markets

Significant Risk from Fluctuating Energy Prices

Today, generators are exposed more to fluctuating energy market prices with the establishment of independent system operators (ISOs), throughout and at renewal of revenue contracts.

Power Generation Industry: Growth Strategy

Predicting Retail Power Prices

Assuming trends in fossil fuel production and pricing markets, can we reasonably predict retail electricity prices by month and state?

Being able to predict changes in electricity prices relative to changes in fossil fuel markets improves strategic visibility for regional acquisitions and could improve risk management by quantifying incoming market impacts from upstream events and improving calibration of hedging portfolios.

Fossil-Fuel-Based Power Prices

Can retail power prices be reasonably determined based only on the cost inputs of fossil fuels and not any renewable power or nuclear generation costs?

If the modeled predictions are reasonably accurate, we can observe market spark margin efficiencies and conclude renewable power companies need to guide their merchant power price strategy based on expected long-term fossil-fuel markets.

Analytical Process

Predicting Electricity Prices

- Electricity prices were predictable from fossil fuel statistics, which represent a source of volatility and risk, and power generation volume, which is related to infrastructure that is easier to forecast.
- The supervised machine learning models were applied to the entire fifty-state data set with the location field removed to start with a generalized approach.
- Data of 6,160 observations in the training set and 1,541 observations in the test set.

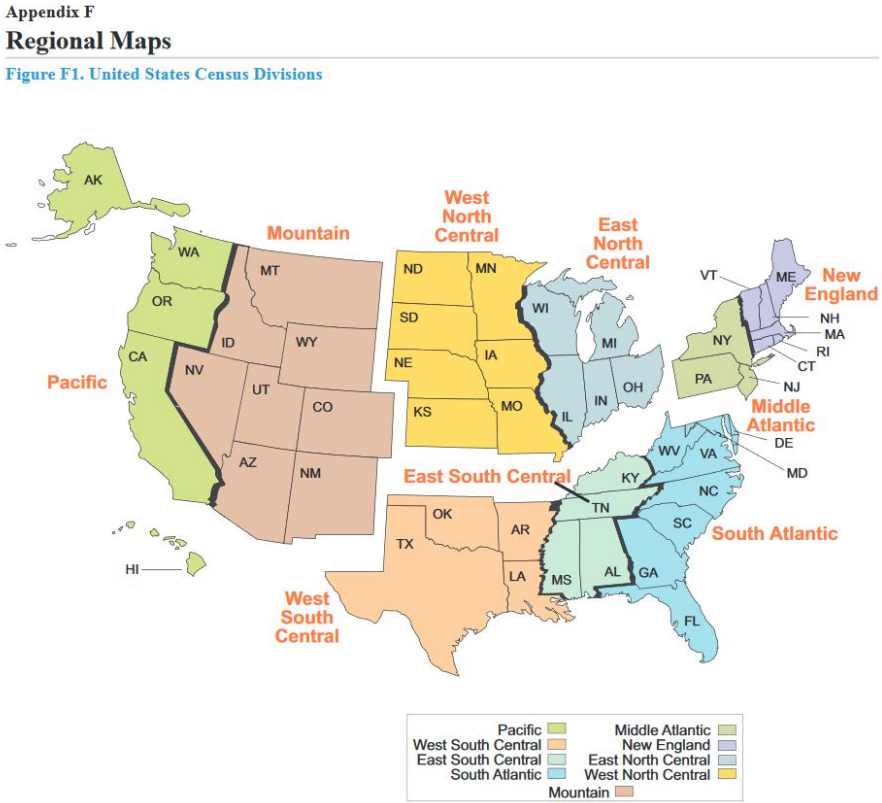
Fossil Fuel Costs Drive Power Prices

- This implies that power generation companies needs to focus risk management data analysis on fossil fuel markets and translate results with an engineering model of thermal efficiency to accurately predict power prices.
- Accurately forecasting energy prices boosts strategic positioning (i.e. if this nuclear plant shuts down, what will power prices do for a time duration) and risk management by calibrating hedging and swap portfolios.

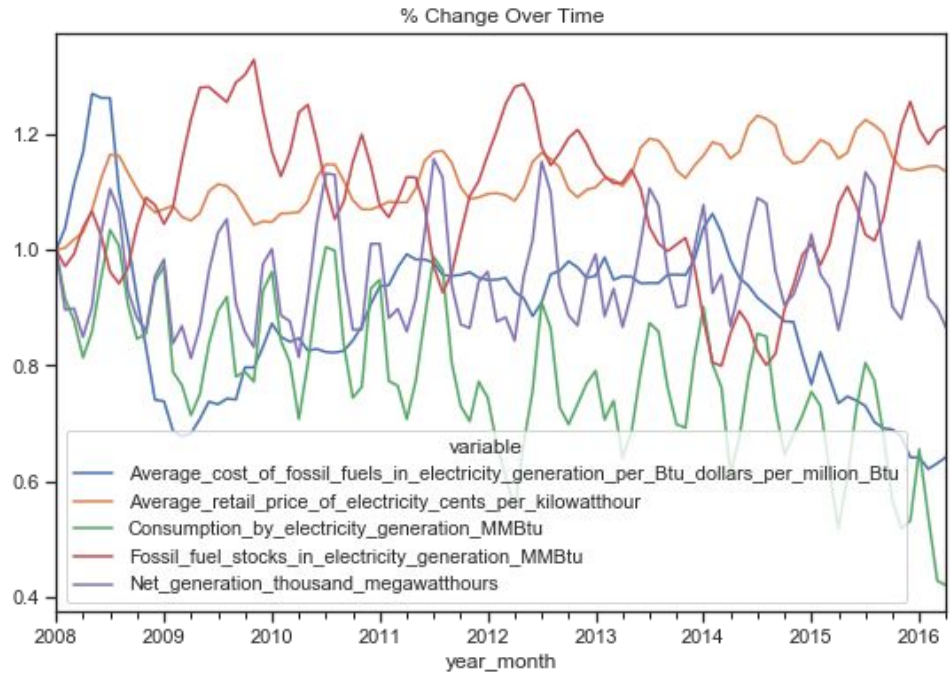
Data Reviewed - Monthly Retail Electricity Prices

Energy Information Administration

- The data consists of monthly average retail electricity prices in kilowatt-hours (kWh) from January 2008 to July 2020, and is broken down by state location and industry sector.
- The data was downloaded from the Energy Information Administration (EIA) website at https://www.eia.gov/electricity/data/browser/.
- Contains 7,701 observations for all 50 states within the United States for twenty-two (22) variables including: target electricity prices, average cost of fossil fuels, generation, fossil fuel stocks, and fuel consumption volumes.



Exploratory Data Analysis (EDA) and Data Cleaning



Comparing Variables' % Chg.

- From a simple percentage change graph, once you remove seasonality out, a visual inspection indicates: Generation has been flat.
- Consumption of fossil fuels has been decreasing overall, with decreased inventory in 2014.
- The average cost of fossil fuels has been decreasing in-line with consumption (a result of demand).

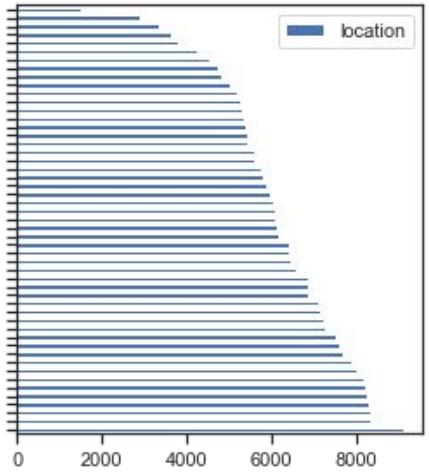
The gradual increase in electricity prices generally tend to appear inflationary and track coal prices, as coal is such a significant portion of the overall power grid.

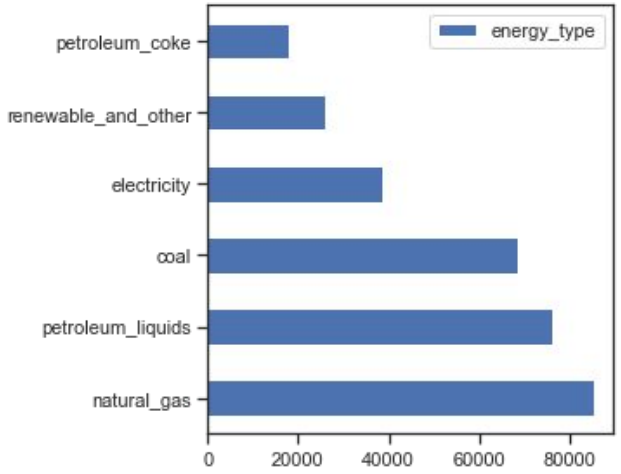
EDA and Data Cleaning

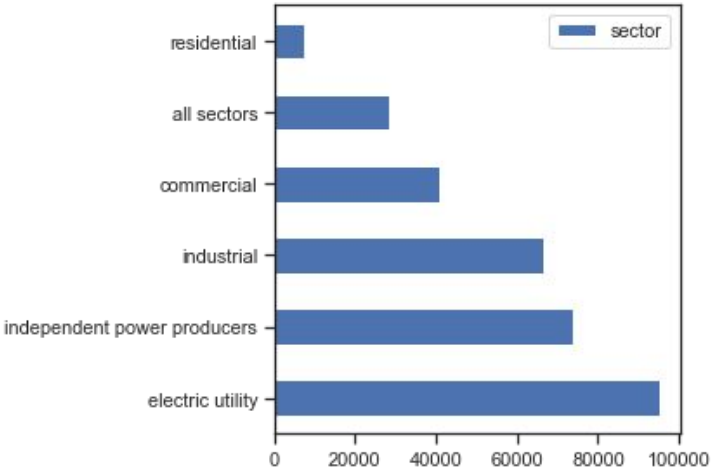
Data Structure: Categorical Groups

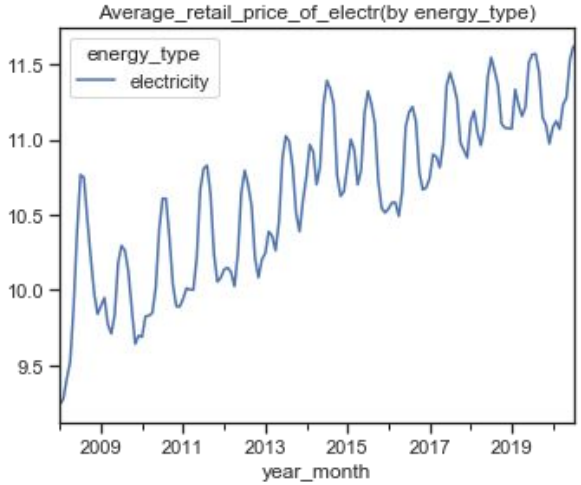
Sector Categories Vary By Variable

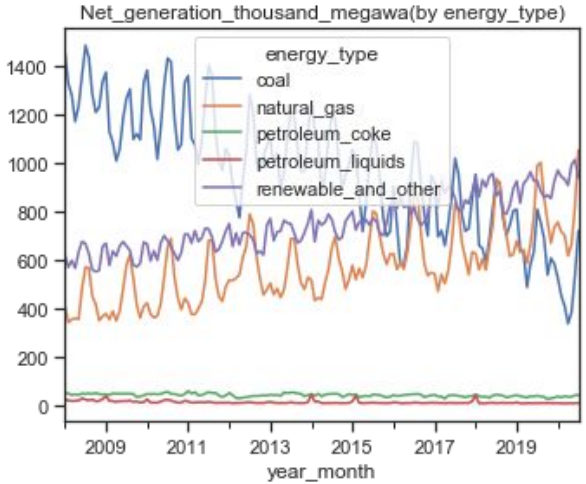
The sector category was almost unique among different variables and did not contain much useful information, therefore the sector level of data was removed and an average was used for each location, energy type, and month.











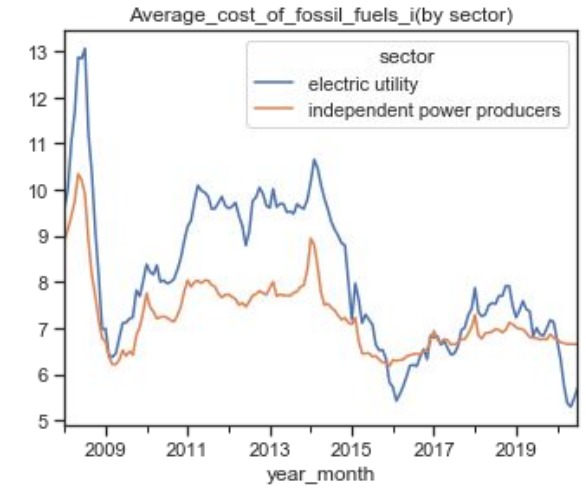
Energy Prices and Generation

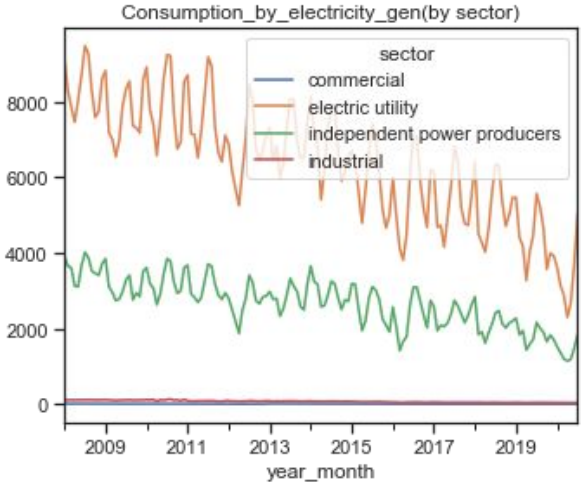
We can see that the average retail price of electricity tends to move seasonally throughout the year by the monthly data. If we had daily data, we would see the variation between on-peak and off-peak hours, but that is beyond the scope of this project.

Coal generation has been decreasing to about half of its level a decade ago, while being replaced with natural gas and renewable energy sources.

Natural gas, petroleum coke, and petroleum liquids fuel prices have trended downward over the past five years as the natural gas glut from fracking continues to maintain negative pricing pressure across the markets.

EDA and Data Cleaning





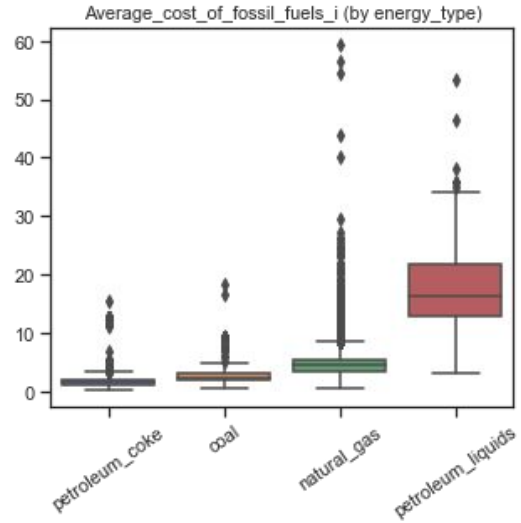
Fossil Fuel Cost Over Time

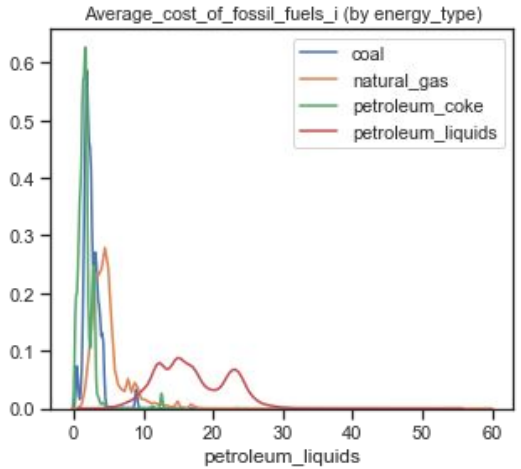
The average cost of fuel between utilities and independent power generators generally shows that private companies are more efficient at maintaining low costs in their operational processes and procurement by a noticeable margin when compared to quasi-governmental utility generators which are heavily regulated.

Looking at consumption of fuels for electricity, utilities are a much larger portion of the market than independent power producers but still cannot garner pricing economies of scale in fuel costs comparatively.

Independent power generators have been steadily increasing their portion of total consumption over the past decade and are now similar to utility companies.

EDA and Data Cleaning



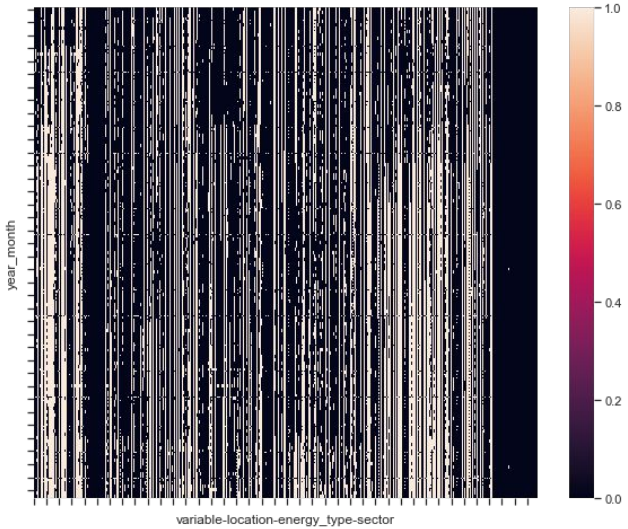


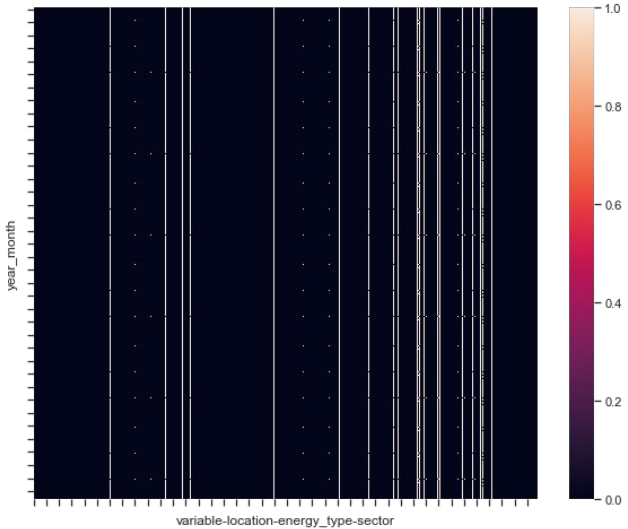
Cost of Fossil Fuels By Type

Petroleum liquids have the largest variation and range in prices per Btu compared to natural gas, petroleum coke, and coal. Petroleum liquids are also the most expensive per Btu, however, petroleum coke is the most inexpensive, high-energy-value fuel in the market and comes off as a bi-product of the petroleum refining process.

The total data points on petroleum liquids are lower than other energy types, so reliability could be slightly compromised with a smaller data set and a comparatively smaller grouping in the dataset

EDA and Data Cleaning

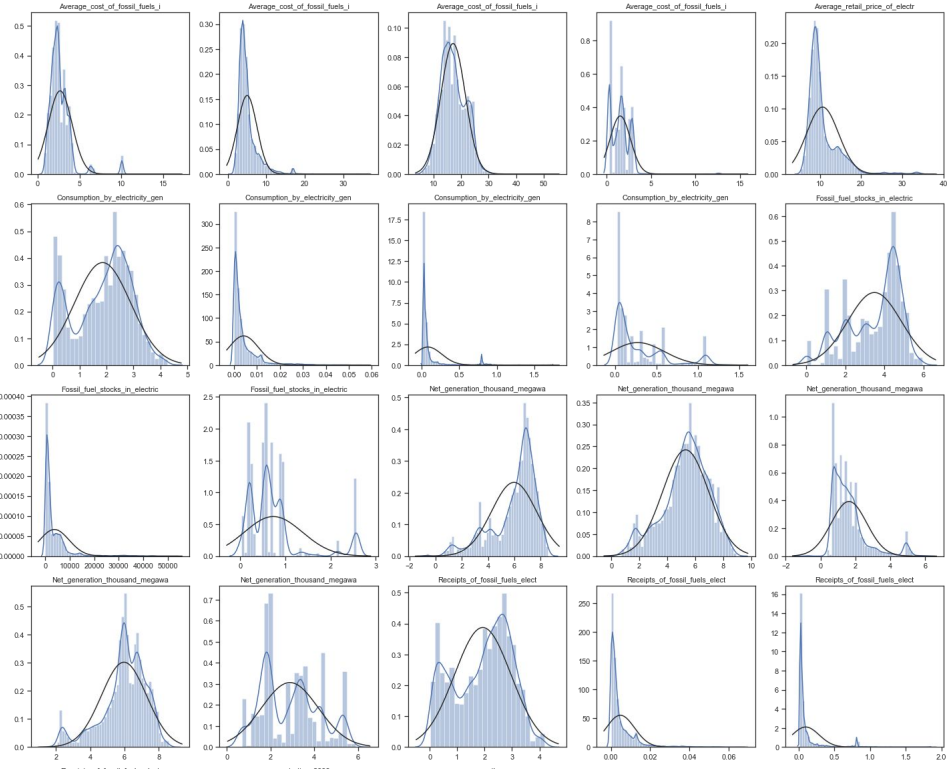




Data Cleaning: Filling

Many of the variables were partially full, or only had sporadic data over the time period. The empty data points were backfilled and forward filled after first breaking them down into groups by location, energy type, sector, and month.

Some data points were not able to be reliably filled, so those columns were dropped (see in the lower image as white).



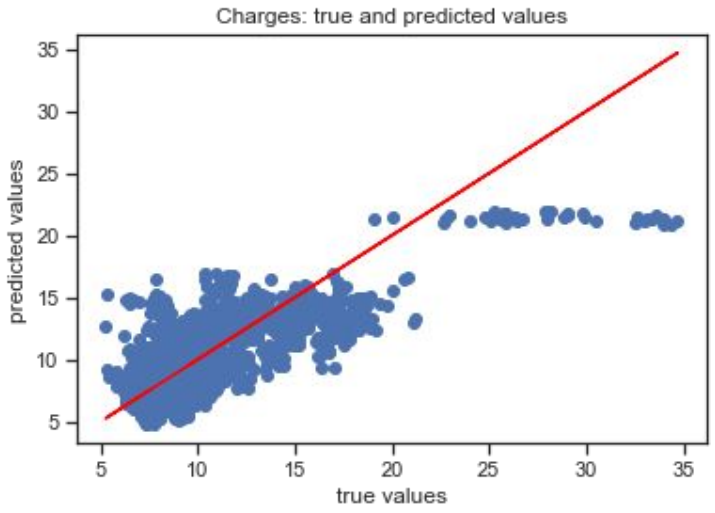
Data Cleaning: Transforming

All non-chronological, continuous feature variables were log transformed to obtain more normal distributions of data. The downside to this is adding a step where the log transformation needs to be reversed in order to understand the units the model coefficients are running off of as they related to reality.

The average retail price of electricity does not have a normal distribution, and appears to be right-skewed.

The fossil fuel variables were combined into homogeneous units by converting all units into MMBtus.

Supervised Learning Results

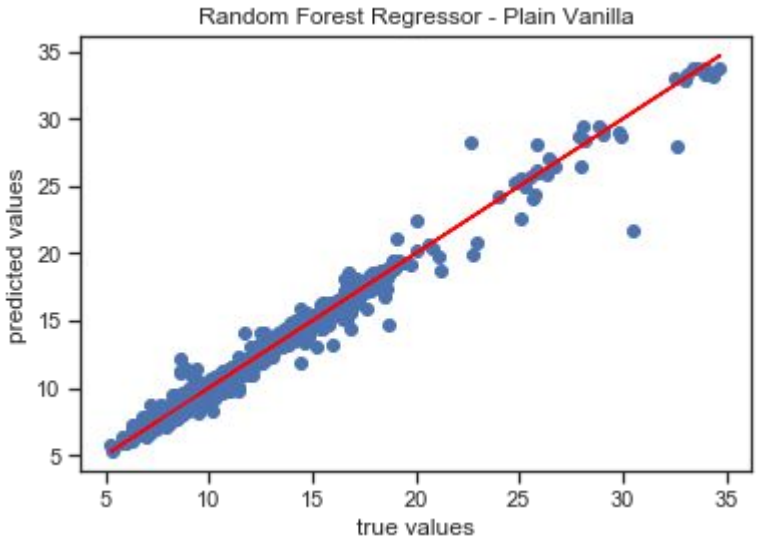


Linear Regression +

- Mean absolute error of the prediction is: 1.9
- Mean squared error of the prediction is: 7.0
- Root mean squared error of the prediction is: 2.6
- Mean absolute percentage error of the prediction is: 18.2
- SK Learn Linear Regression - Adjusted R-squared value: 0.57 with RMSE of 2.6.

index	r2_training	r2_test	avg_abs_err	avg_sqrd_err	root_mean_sqrd_err	avg_abs_pct_err	est_regularization_alpha
Ridge Regression with CV	0.548	0.562	1.957	7.142	2.672	18.486	0.100
Lasso Regression with CV	0.497	0.503	2.018	8.112	2.848	18.823	0.100
ElasticNet Regression with CV	0.522	0.528	1.985	7.706	2.776	18.471	0.100

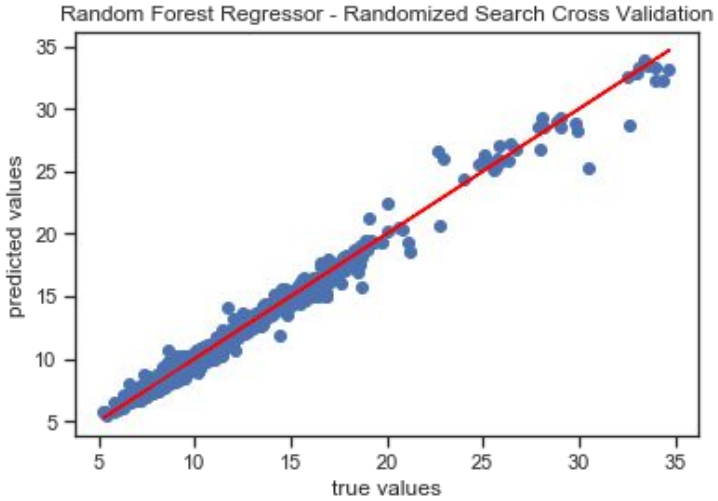
Supervised Learning Results

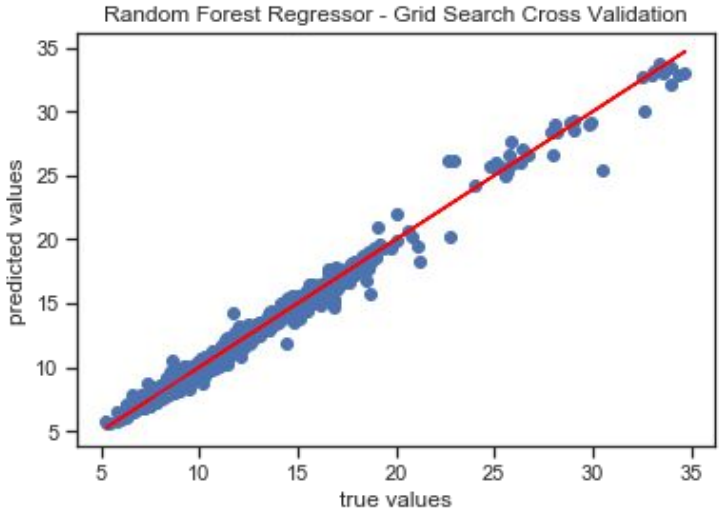


Random Forest Regression

- Adjusted R-value: 0.98
- Mean absolute error of the prediction is: $0.32
- Mean squared error of the prediction is: $0.32
- Root mean squared error of the prediction is: $0.56
- Mean absolute percentage error of the prediction is: 39.3%

Supervised Learning Results



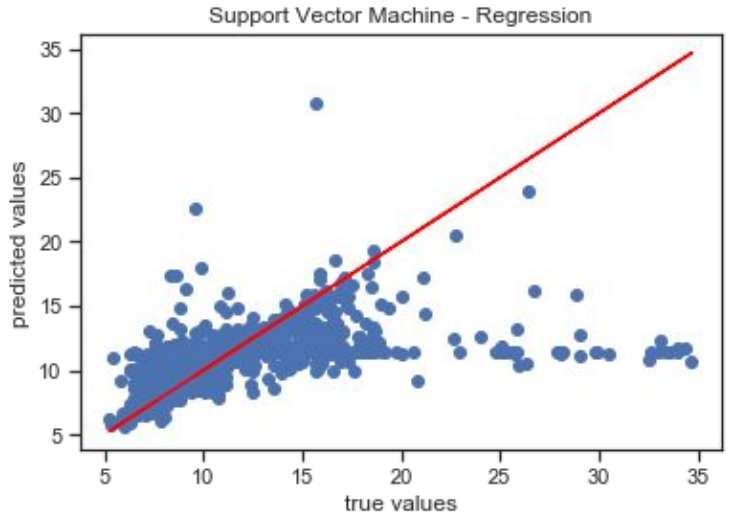


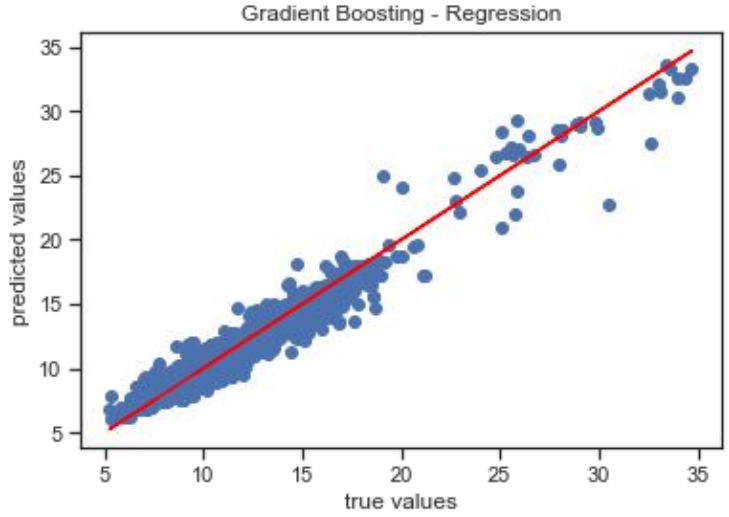
Random Forest Regression with Hyperparameter Tuning

- Adjusted R-value: 0.99
- Mean absolute error of the prediction is: $0.29
- Mean squared error of the prediction is: $0.22
- Root mean squared error of the prediction is: $0.47
- Mean absolute percentage error of the prediction is: 39.4%

- Adjusted R-value: 0.99
- Mean absolute error of the prediction is: $0.29
- Mean squared error of the prediction is: $0.20
- Root mean squared error of the prediction is: $0.45
- Mean absolute percentage error of the prediction is: 39.5%

Supervised Learning Results





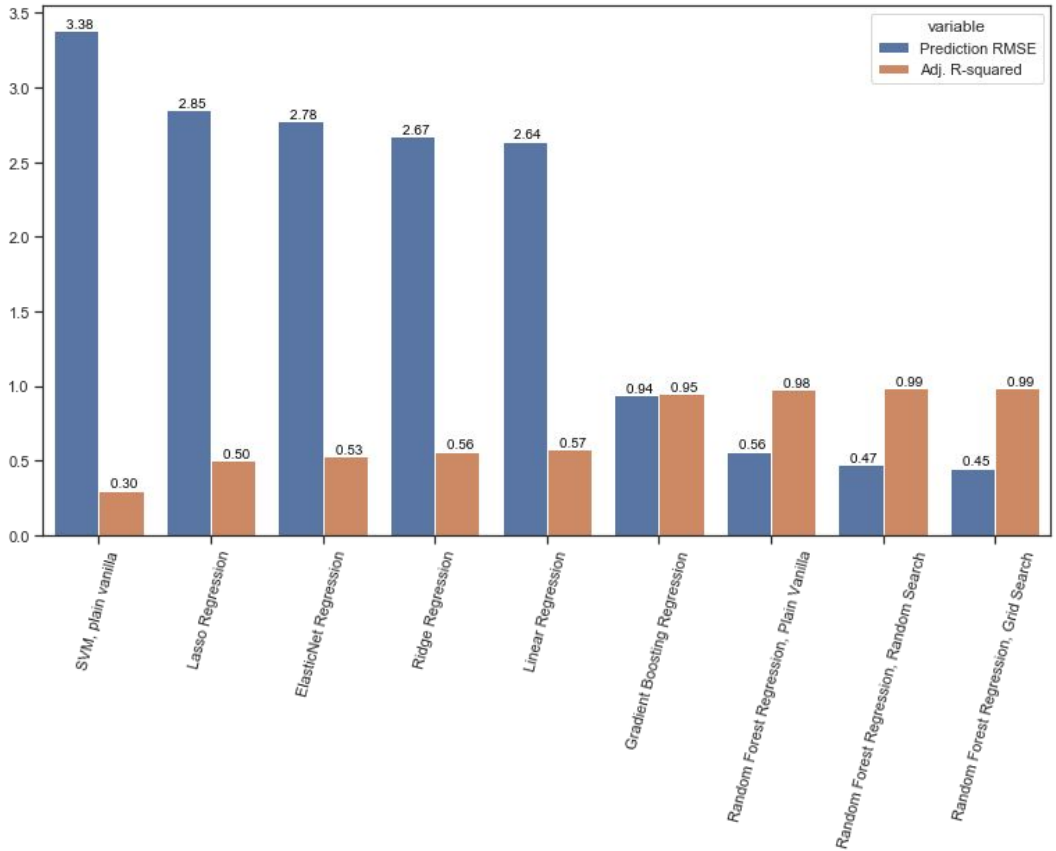
Support Vector Machine (SVM) Regression and Gradient Boosting Regression

- Adjusted R-value: 0.30
- Mean absolute error of the prediction is: 1.8
- Mean squared error of the prediction is: 11.4
- Root mean squared error of the prediction is: 3.4
- Mean absolute percentage error of the prediction is: 32.0%

- Adjusted R-value: 0.95
- Mean absolute error of the prediction is: 0.7
- Mean squared error of the prediction is: 0.9
- Root mean squared error of the prediction is: 0.9
- Mean absolute percentage error of the prediction is: 38.4%

Supervised Learning Results

Supervised Learning Results - Random Forest



Conclusions

The Random Forest Regression was the best performing model, particularly when hyperparameter tuning with cross validation methods.

	Prediction RMSE	Adj. R-squared
SVM, plain vanilla	3.378	0.301
Lasso Regression	2.848	0.503
ElasticNet Regression	2.776	0.528
Ridge Regression	2.672	0.562
Linear Regression	2.641	0.572
Gradient Boosting Regression	0.940	0.946
Random Forest Regression, Plain Vanilla	0.562	0.981
Random Forest Regression, Random Search	0.470	0.986
Random Forest Regression, Grid Search	0.450	0.988

Questions and Discussion

Thank You