Adam Yankelevits, Aaron Drew, Eric Yehl, Arbaaz Shakir, Priyadarsini Subramanian, John Stuart

Final Write Up: Electricity Price Prediction

Utility residential electricity prices have risen steadily in the last decade. According to the Energy Information Administration, residential electricity rates have increased nationally by around 20% in the last 10 years – from about 10.8¢ per kilowatt-hour (kWh) in 2007 to about 13¢/kWh in 2017 (an increase of more than 0.2¢/year).

This project aims to forecast changes in the price of electricity based on the factors we have identified as influencing the price, giving policy makers, investors and venture capitalists a tool to guide their decisions for energy and other infrastructure projects.

Forecasting predictions in the future is always a challenge. Compounding this, is the fact that there are so many factors that could influence the price of electricity. There are the common factors that influence electricity prices like demand, consumption and cost of fuel for generators. Additional factors that may dictate the electricity prices include capacities of electric utilities, supply of alternate energy sources and new technologies which will reshape the energy landscape such as electric vehicles and lithium ion batteries. And of course, electricity prices vary by time and location.

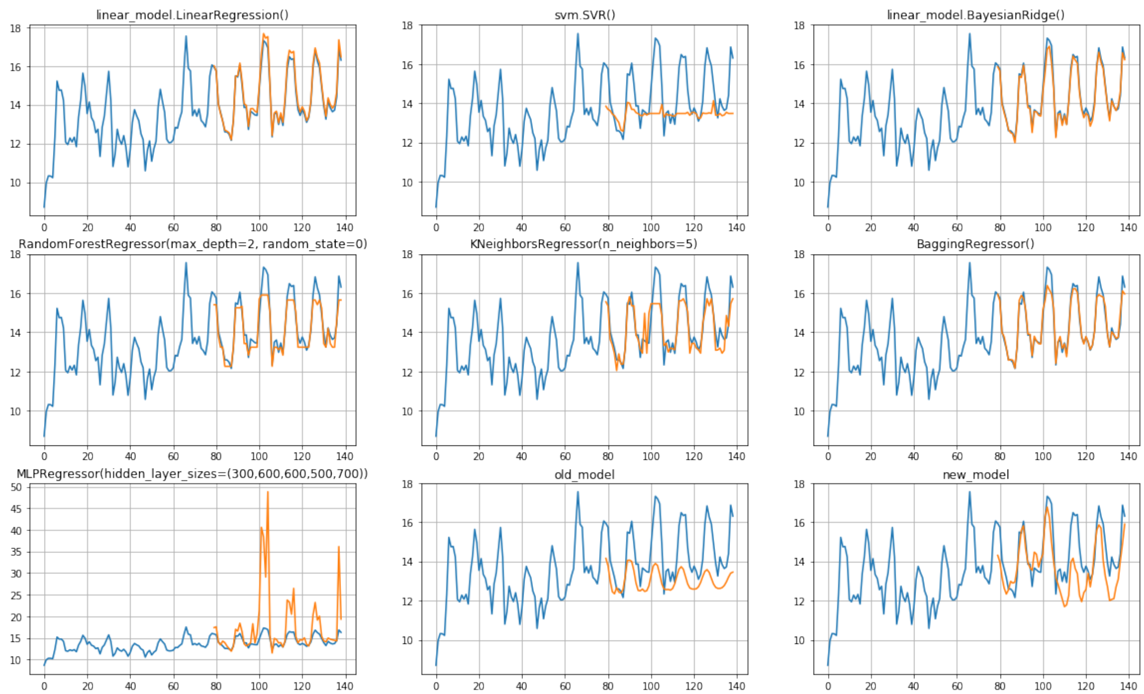
Another problem we found was that any model we were going to look at predicts the price of electricity for a given month, assuming that we have full information for that month. However, in reality, one would not have data for any month until the end of the month, and by that time, prices would already be predicted. So, we wanted to come up with a way to take into account data from the past and see how it correlates to a current month's electricity price.

Therefore, predicting the electricity prices for the next 5 years with high level of accuracy is a huge challenge based on current models. Our team is using the best combination of factors (ie. features) and statistical models to propose a high-confidence price prediction model for the next 5-10 years to serve as the backbone of user-facing applications.

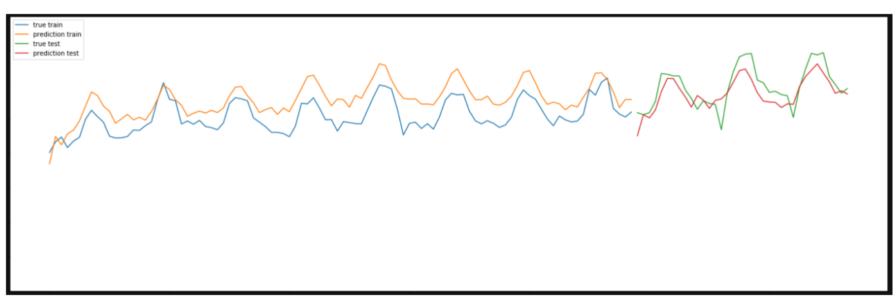
Our solution was multifaceted. First, we wanted to collect more data and have a larger set of features. In addition to last semester's data, we added capacity data, EV data, and market stock prices for electricity utilities. With this larger data set, we performed feature construction—building non-linearity into a linear model. We took sums, products, quotients, among different transformation of the data. Lastly, we rather than only looking at features and electricity prices in the same month, we looked at the value of features in the past (in one-month increments) to see if there is a time delay for the effect on energy prices. This gave us a very large set of features—up to around 10,000 at times—for which the runtime and computing power necessary would be too great to work with. Then depending on the model we used, we came up with different ways to cut down on the features. One way was a simple Pearson correlation, finding features that

are highly correlated with price, and poorly correlated with one another. Another solution was recursive feature elimination followed by principle component analysis to reduce the dimensionality and significantly cut runtime. This also prevents overfitting as we lose a slight amount of information, but it means we cannot get 100% training accuracy.

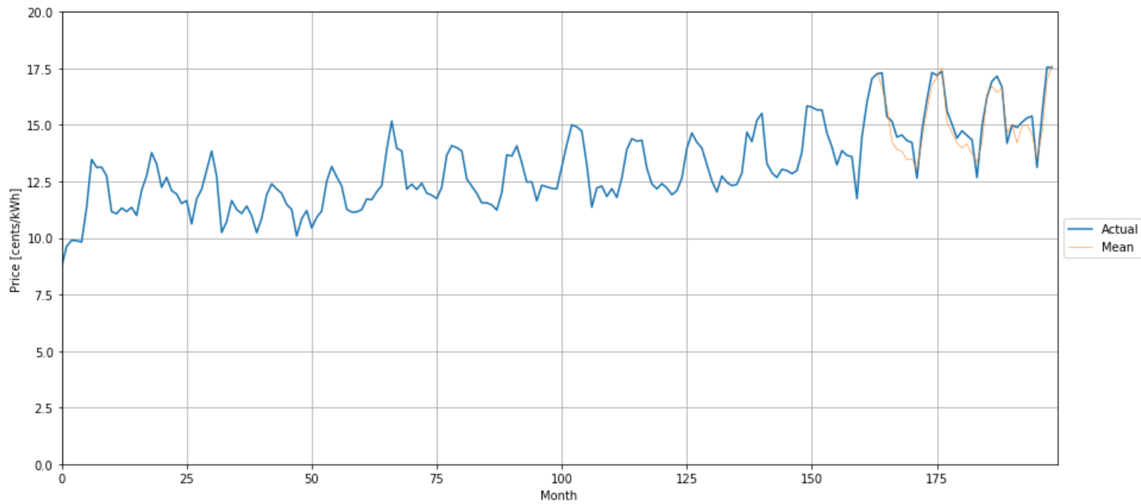
After going through an intense process of feature selection, we built different models to test, hoping to find the most accurate models. First, we built a docket of standard models that include linear regression, Bayesian ridge regression, SVR, KNN, random forest regression, bagging, multi-layer perception, and last semester's SARIMA model. We ran different sets of data on each of the different models, and the output was the RSMEs of each model/data-set pairing.



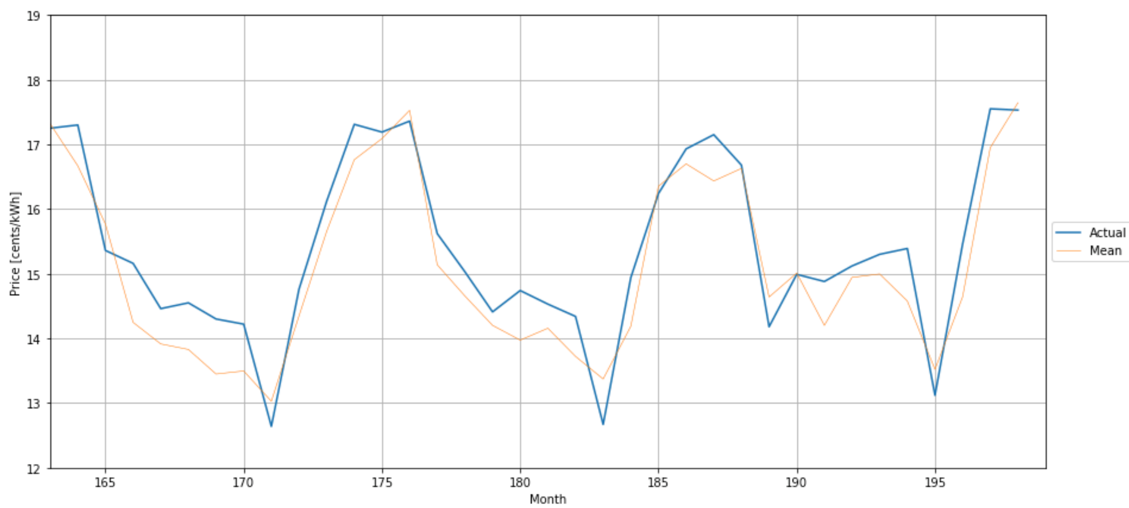
A second approach we took to solving this problem was to build an LSTM network, that first decides which features to throw away or keep, then decides which information should be stored in each cell. We have included a sigmoid layer and a tanh layer that decide which values to update and create a vector of new candidate values, respectively. The final gate updates the old cell state into a new cell state that then predicts the prices in a time series. We implemented a GRU that merges the cell states and hidden states because it is less computationally intensive without sacrificing much performance.



The third approach we took was using a LASSO regression on the features we pulled from using the time delays and feature construction that left us with the dataset of about 10,000 features. For every month that we want to predict in the future, it trains an independent LASSO regression using past data as present features. Put another way, given any month we want to predict prices for, we train a model that looks at data points in the past. This model also optimizes four hyperparameters. The first is the look-back window, how far in the past we are looking at data to train the model on price; alpha, the regularization coefficient; a special type of interpolation that gives heavier weights to more recent data points; and the proportion of the data set to train and test on.



Mean RMSE: 0.531 cents



With these methods, we have not only been able to predict energy prices into the future, but we have built a pipeline that efficiently can take large sets of data and work on finding the most effective features as well as the most effective models. As our work progressed on this project, without losing focus on the goal of electricity price prediction, we became more and more interested in building this pipeline that can serve as a blueprint for other data science problems. We hope that this can be used to predict prices or any other time-series data in a variety of industries in the future.