Long-Term Electricity Price Prediction

Members:
Kun Yang, Shikhar Verma, Nami Saghaei, Randell Espina, Maria Sheila Ramos, Ashkan Yousefi
Mentors:
Antonio Vitti, CFO formerly with Merchant Atlas, Inc.
Dr. Steven Gustafson, Chief Scientist at Maana

Introduction

Prior to 2008, investments in large power-plant and energy infrastructure projects were largely led by big domestic banks. Since the 2008 market crash, however, investments in these assets have run dry, largely because banks have been dissuaded from investing in long-term and risky energy assets. Nevertheless, energy in the United States has proven to be an innovative and profitable sector in recent years, and newly available capital and opportunities leave investors eager to get involved. This has left a once-in-a-lifetime hole in the market in which there is significant demand for innovative energy projects with great potential for return, and eager investors, but a disconnected investment scene both domestically and abroad, mostly due to lack of information and uncertainty about the projected financial performance of these projects.

Our project aims to use newly available data and novel data analysis techniques to produce a reliable prediction for the monthly retail price of electricity 3-5 years in advance. This analysis platform would serve as a long-needed tool for every-day investors and analysts to conduct state of the art risk analysis/investment evaluation in this space. As a solution that outperforms the state of the art, this tool would be a huge attraction for energy investors, who have been reluctant to invest in long-term projects because projections for their investments are not typically backed by a non-trivial data-driven forecasting model, and instead some arbitrary group's intuition.

The Underlying Problem

The low-level problem here is that there exists a massive gap between the analysis currently being conducted in the field, and the analysis that could potentially be done given the abundance of data readily produced by the sector and the emergence of novel machine learning techniques in the last few years. We hypothesize that, by taking advantage of these new developments, we can offer analysis that can bridge this gap, provide long-term insights that have otherwise been impossible to obtain, and guide investors in viable long-term energy investments.

Process and Solution

Data Collection and Cleaning

Finding and cleaning complete and reliable data was the toughest part of this semester's work. For most of the semester, we had disorganized and scattered data which could not be readily converted into a standard format. We tried understanding the collected data and attempted developing more robust models. During the last part of the semester, we found an extremely reliable source of public information in the Department of Energy's OpenData database, which we were able to query, export to CSV, and run our analysis on, after cleaning and merging the data with other data that we had gathered throughout the semester. Finally, towards the end of the semester, we were able to obtain 78 features and made our electricity price prediction.

Correlation Analysis

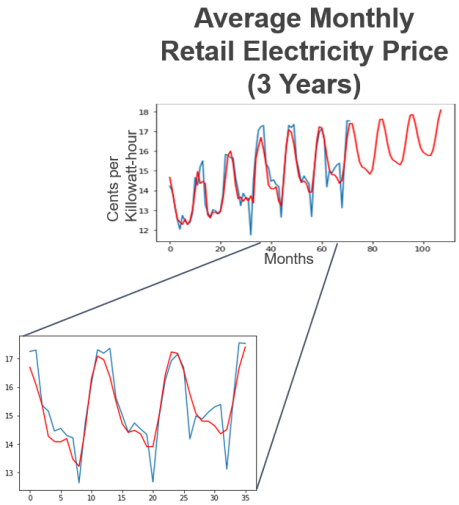
We ran some simple correlation analyses on the cleaned data, and were appalled by the number of seemingly unrelated factors that highly correlated with the average price of electricity. We were also particularly surprised by the huge correlation of renewable consumption and production with the retail price of electricity. We hypothesize that this is because renewables are not well-established like Oil, Coal, or Gas, all of which have thousands of non-quantifiable factors that come into play when determining price. Instead, it turns out that the main driver behind production and consumption of renewable energy is still people; when electricity is cheap, there is no incentive for people to incur the heavy startup costs of buying $10,000+ solar solutions. And as electricity gets more expensive and solar adoption becomes more affordable, people feel more comfortable switching to renewables because they save money.

This was a huge find for us. Because of its relative renewable-friendliness, California's energy prices can likely be more predictable than the rest of the country. We, of course, took advantage of this insight, and more than a third of our most highly-correlated features ended up being renewable-related.

Prediction with Time Series

After trying several models, we settled on some not-so-sophisticated Time Series analysis, using Autoregressive Integration with Moving Average (ARIMA) via the autoregression model in the Statsmodels Python Package (statsmodels.tsa.ar_model.AR), which was able to pick up extremely well on the seasonality of the data. Here we learned that the monthly price of electricity, when controlled for location, is very seasonally predictable, with repeating trends each season and a predictable increase of average prices occurring each year for recent years. Our 3 year time series predictor was able to capture complex trends to a surprising degree and

performed extremely well, with an average monthly error of less than half a cent on the 2014-2017 test set, and an accuracy within a cent of the true price in ~90% of the months. Our 5 year time series predictor, however, overfitted to trends coming up to 2012 and projected higher prices than actual. We believe this error to be the result of a shrinking train set of 2001-2012, and the the fact that 2013-2014 were particularly trend-setting years. Without data from those years, the algorithm produced a much less reliable prediction.



MSE (Mean Square Error)=0.38

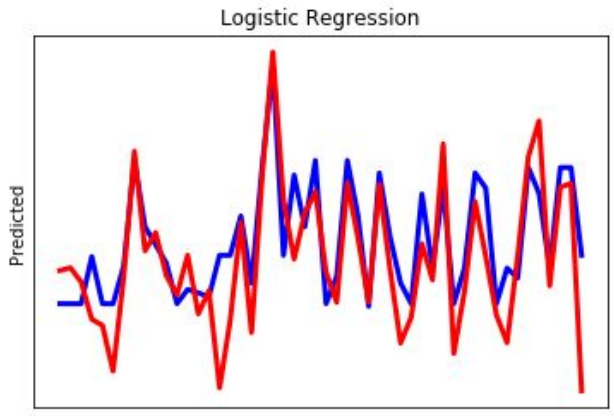
MSE (Mean Square Error)=4.972

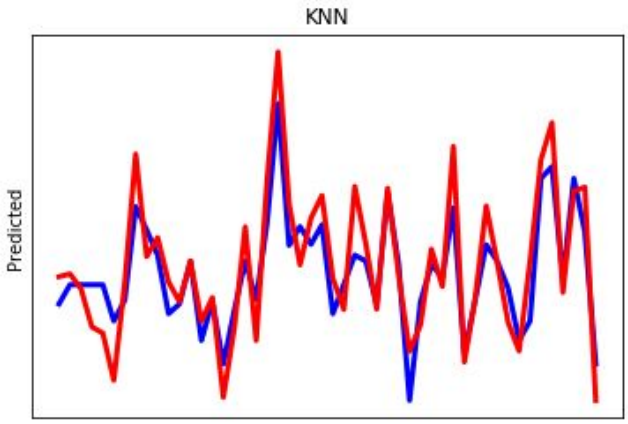
Hand-Picked Features Model

We then wanted to train a model on the hand-picked features with the ultimate goal of beating the 5 year time series predictor. The idea was to:

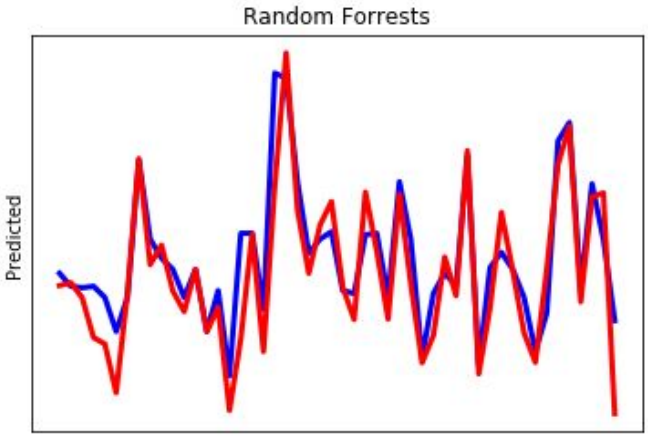
- Use Time Series to predict value of features in the future.
- Train a model with the highest-correlated features.
- Apply coefficients of tuned algorithm as weights to projected features to obtain a projected future retail cost of electricity.

We made great progress on this objective, and compared our top 3 best performing algorithms in the presentation.





This model is still in development, and we hope to improve it in the coming months to beat the 5 year predictor and, hopefully, even the 3 year predictor. We will also investigate some sort of multi-model weighting or a mixture of our Time Series prediction and the learned feature-based prediction. We think that elaborating on the complexity of the model will be our next main step for the coming semester.



Challenges and Other Next Steps

One really hard aspect of this problem (and one of the main reasons nobody has really solved this problem) is that there are factors that simply cannot be predicted. One thing we would like to do in the future, given more time and resources, is to look at a way to provide several perturbance-flexible estimates to be tuned with domain knowledge. So some logical next steps are to figure out a way to quantify perturbances, develop the front end, and investigate how to incorporate domain knowledge into the prediction algorithm.

Contribution of team members

Kun Yang: Collected the data, cleaned and merged data, selected the high correlated feature, worked on the time series prediction.
Nami Saghaei: Cleaned and merged data, conducted correlation analysis, selected highly correlated features, tuned/analyzed time series prediction, summed up the project and made the presentation.
Shikhar Verma: Worked on Hand-Picked Features Model; selected highly-correlated features, trained and analyzed models.
Randell: Collected and cleaned the data, conducted initial correlation analysis.
Maria: Collected and cleaned the data.
Ashkan: Searched the resources and determined the method.

Huge thanks to Ikhlaq, Alex, and Antonio for providing vital guidance throughout this project. We look forward to working with you next semester.

Link to the Project's Github repo: https://github.com/afbholly72/Energy_Price_Predictor.git

Reference

[1] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[2] Jason Brownlee (January, 2th, 2017) Autoregression Models for Time Series Forecasting With Python https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

[3] Statsmodels http://www.statsmodels.org/stable/index.html