Long-Term Hedging of Electricity price Using Machine Learning Techniques and New Feature Discovery

Ashkan Yousefi UC Berkeley

***Abstract*—Predicting the performance of energy commodities has long been a global priority for researchers and investors in the Energy sector. Large projects (often exceeding 1bn USD) are financed with locked in capital from the start, and typically take decades to return. Despite being one of the most important aspects of investment decision making, the prediction methodologies used widely today are not sophisticated enough to provide any hard-quantifiable insights. The new approach was proposed to provide data analytics backed hedging for the performance of energy related commodities using novel feature discovery methods and machine learning tools. In the presented research, a machine learning model was trained to predict the average monthly price of electricity in the next 5 years with focus on the California State. Data points from 2001 to 2017 are collected and 78 data points are considered for various analyses to find the highly-correlated features which could affect the electricity price. An economic case study is undertaken to understand the correlation of the features, and to avoid multicollinearity. In the next step, the features are fed into an S-ARIMA time series prediction algorithm, as well as several feature-based machine learning algorithms. The findings demonstrated promising results for 3 years future price prediction. The analyzes results suggest that 5 year results are attainable.**

Keywords—Machine Learning Algorithms, Clustering, Random Forests, Time Series Analysis, Feature Correlation Analysis, Synthetic Data Augmentation, Hedging, Energy Price Prediction, Commodity Pricing

I. INTRODUCTION

Prior to 2008, investments in large power-plant and energy infrastructure projects were largely led by big banks. Since the 2008 market crash, however, investments in these assets have run dry, and 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 created a gap in the market between the significant demand for innovative energy projects with great potential for return and eager investors. The disconnected investment scene exists, 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 performance of energy commodities, namely the monthly retail price of electricity 3-5 years in advance. This analysis pipeline would serve as a long-needed tool for every-day investors and analysts to conduct state of the art risk analysis in this space. In addition to applications in institutional investments, this methodology could aid everyday economists and researchers in making their own forecasts and creating a noticeable impact on policy makers in energy industry in determining the penetration level for demand response programs for the renewable energy projects.

II. EXISTING WORK

The energy industry is one that produces ample data thanks to the large-scale implementation of sensors, and the general interest of both the companies and the public. As a rather traditional industry, however, it does not tend to promote the use of cutting edge data analysis techniques [1]. Internally, analysis is most often conducted by consulting companies who use simple forecasting tools in Excel, combined with their knowledge and intuition about the industry. Better analysis is done by independent data scientists on public data science platforms like Kaggle [2], but energy projects are not as widely available on the platform as the ample availability of the data may suggest.

The low-level problem here is that there exists a massive gap between the analysis currently being conducted and relied upon 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 simple and powerful machine learning techniques in the last few years. We posit that, by taking advantage of these new developments, we can offer analysis that can bridge this gap, provide researchers with long-term insights, and guide investors in viable long-term energy investments. The use of advanced tools can help the investors to achieve more clarity for their investment decision to plan for the next five years.

III. RESEARCH AND TECHNOLOGY APPROACH (ANALYSIS AND METHODOLOGY)

A. Data Collection and Cleaning

Data from the US Energy Information Administration's OpenData database [3] (one of the earliest completely open government data portals) was utilized for this project. Querying for the data in CSV format through their API, minor cleaning was conducted (column names, local average for missing values when reasonable), and datatables were merged to get consistent data from 78 different commodities and services for every month between January 2001 and August 2017.

B. Feature Correlation Analysis

Correlation analysis was conducted via generation of a Pearson correlation matrix on the cleaned data using the Pandas Python package [4] in order to find the data that was most highly correlated with the retail price of electricity. We summarize highly correlated features, and present interesting insights in *Findings*.

C. Time Series Prediction

An Autoregressive Integration with Moving Average [5] model was trained via the autoregression model in the Statsmodels Python Package (statsmodels.tsa.ar_model.AR [6]) to produce two end models;

1. Model aiming to predict prices 3 years in advance - train 2001-2014, test 2014-2017, predict 2017-2020.
2. Model aiming to predict prices 5 years in advance - train 2001-2012, test 2012-2017, predict 2017-2022.

Performance of these models is analyzed and compared in *Findings*.

D. Feature-based Model Prediction

Following the promising performance of the S-ARIMA model in capturing seasonality of energy data, we investigated the use a simple machine learning model on the ARIMA-projected values of features in order to predict electricity price. Several models were trained on the hand-picked features with the goal of achieving a lower mean squared error than the 5-year time series predictor. The methodology proceeded as follows:

- Used Time Series analysis to make predictions for values of highest-correlated features in the future.
- Trained several models with true values of highest-correlated features in the present.
- Applied coefficients of tuned algorithm as weights to projected values of highest-correlated features to obtain a projected future retail cost of electricity.

We discuss and compare our best performing algorithms in *Findings*.

IV. FINDINGS

A. Data Collection and Cleaning

Despite the ample availability of public energy data on the web, obtaining consistent and reliable energy price data for a controlled geographic region was a major challenge. Data was gathered from tens of sources and merged, then compared with that published by each hub belonging to energy companies in Southern California. Inconsistencies were found in almost every data table obtained from merging data from various sources. Finally, reliable energy data was found in the Energy Information Administration's OpenData database [3], and their open-source data was used exclusively in order to avoid inconsistencies in measurement techniques, scope, and reliability.

B. Feature Correlation Analysis

Figure 1 demonstrates a correlation matrix detailing the correlation of all the 78 features in the data set.

The following factors were extracted for use as input features in the analysis. They were chosen due to their high correlation with electricity price. as well as reliability. Numbers indicate correlation of feature with Average Monthly Retail Electricity Price:

- Natural Gas Consumed, Electric Power Sector (.85)
- Electricity Generated by Petroleum Coke (-0.66)
- Net Electricity Generation ,Commercial(.81)
- Net Electricity Imports (.75)
- Natural Gas Consumed by Industrial Sector (.74)
- GDP (.7)
- Renewables (.77)
    - Solar Consumption (.71)
    - Geothermal Consumption (.65)
    - Wind Consumption (.62)
    - Biofuels Production (.72)

We were surprised by the number of seemingly unrelated factors that were highly correlated with the average price of electricity. Perhaps most surprising was the high correlation of the consumption and production of renewable energy features to the retail price of electricity. Solar, Geothermal, Wind, and Biofuel Consumption and overall Renewable Generation were all heavily correlated with the retail price of electricity.

Our findings seem to agree that there exists a causal relationship explaining the influence of the price of electricity on the adoption of renewables. As detailed in Figure 2, the correlation plots produced by Natural Gas Consumption and GDP (both highly correlated with electricity prices) are scatter plots with a clear positive correlation (a positive trend is shown in the plots). However, the correlation plot produced by Solar Consumption and Production, Wind Consumption, and the production and consumption of Renewables as a whole (Figure 3) indicate a curve wherein the increasing price of electricity seems to cause a reactionary increase in the adoption of renewable energy.

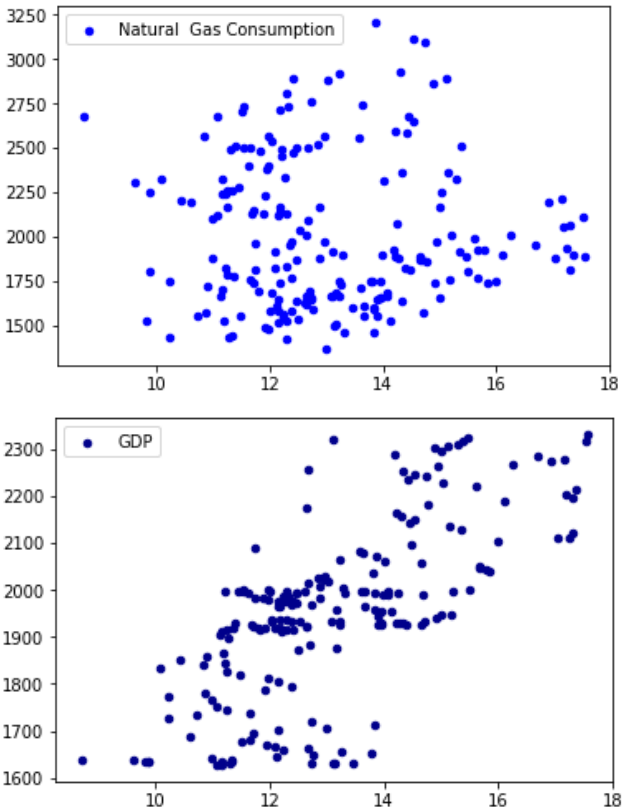


Fig 2 Natural Gas Consumption and GDP vs. Electricity Price

Unlike the well-established fossil fuels industry, where a number of non-quantitative factors (political climate, foreign trade, etc) unpredictably influence the price of electricity, the main driver behind trends in production and consumption of renewable energy is demand; when electricity is cheap, there is little incentive for people to incur the heavy overhead costs of buying renewable solutions such as solar panels. And, as electricity gets more expensive and renewables become more affordable, people feel increasingly comfortable switching due to more promising savings. Due to its relative renewable-friendliness, California's energy prices would likely be much more predictable. Conclusively, more than a third of our most highly-correlated features connected to renewables.

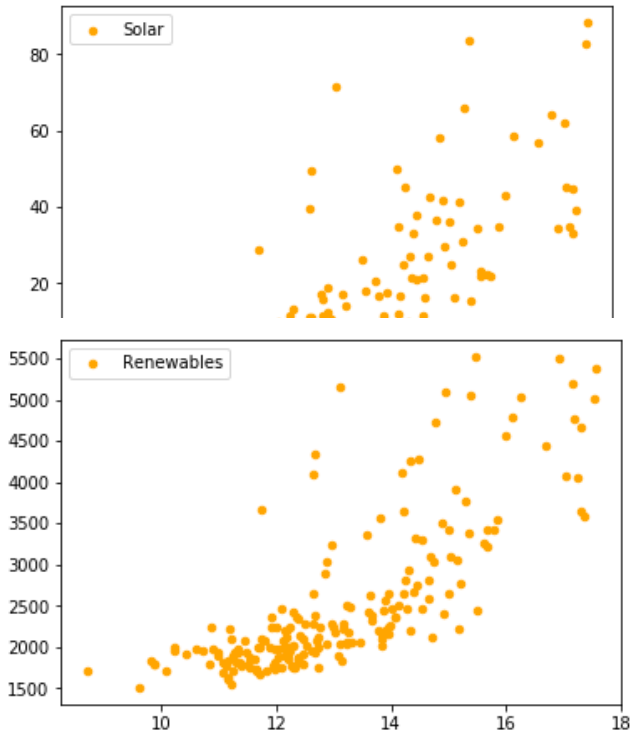
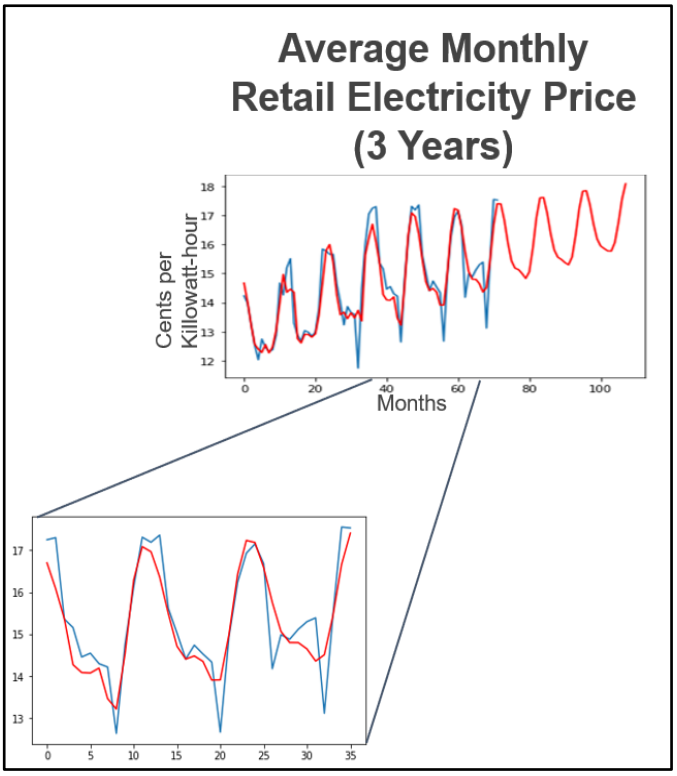


Fig. 3 Wind, Solar, Renewables Consumption vs. Electricity Price

C. Time Series Prediction

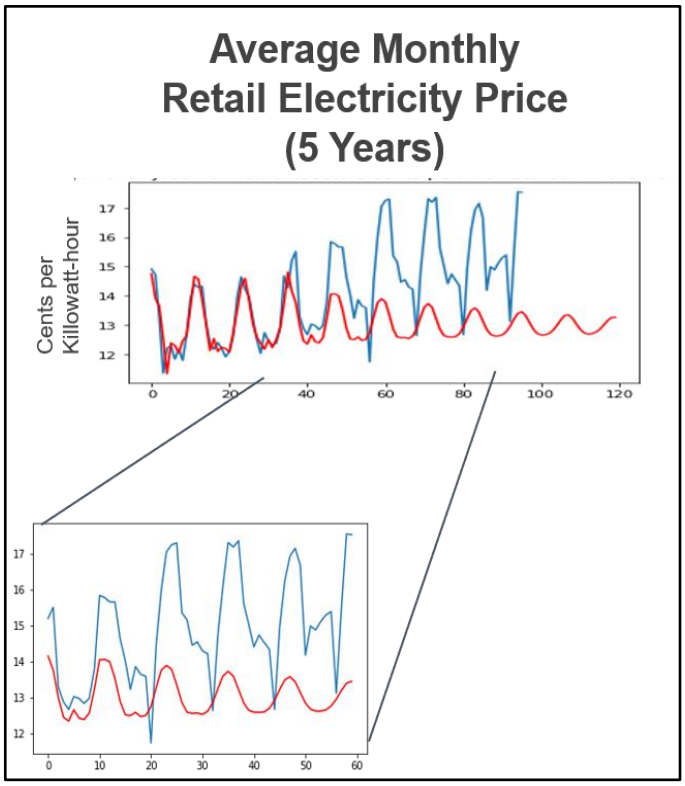
Our analysis indicates that the monthly price of electricity, when controlled for region, has a high seasonality with a relatively predictable increase of average prices occurring each year for the recent years. Our 3-year ARIMA time series predictor was able to capture complex trends to a great degree 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 (Figure 4). Our 5-year ARIMA predictor, however, overfitted to trends coming up to 2012 and projected prices to be consistently higher than the actual. We believe this error to be the result of a diminished train set of 2001-2012, and the the fact that 2013-2014 were particularly trendsetting years in

renewable and electricity pricing; with no data during those years, the algorithm produced a less reliable prediction. Figure 5 compares the monthly values as predicted by the model with actual values, and illustrates this overfitting.



MSE (Mean Square Error)=**0.38**

Fig. 4 Performance of 3-Year ARIMA Predictor

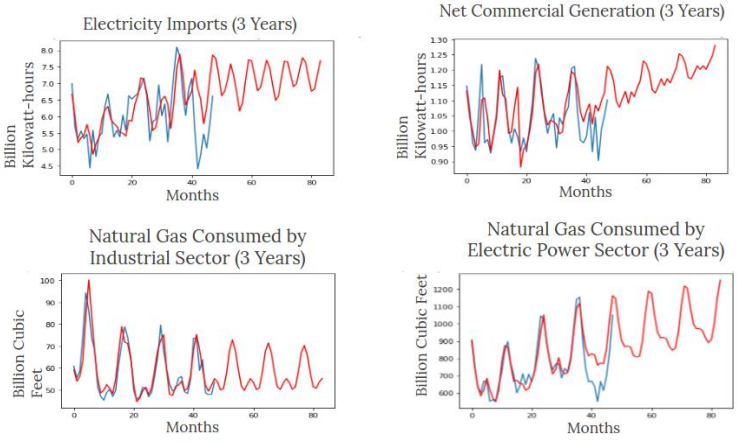


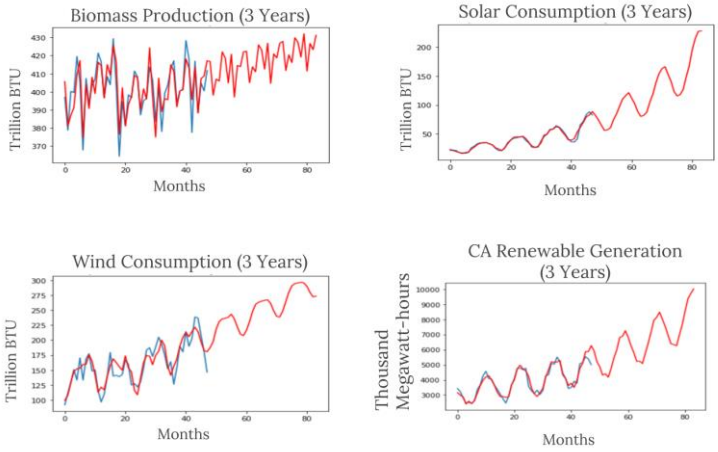
MSE (Mean Square Error)=**4.972**

Fig. 5 Performance of 5-Year ARIMA Predictor

D. Feature-based Model Prediction

After preprocessing, S-ARIMA was used to synthetically augment the data from the 11 selected features from 2001-2014 in order to create projected values to compare with the 2014-2017 test set. The plots below present monthly





projected values in 2014-2017 overlayed with the true values of 8 of 11 (most reliable) features, in addition to further projection into the future.

Upon performing linear regression on the augmented dataset, we obtained MSE = 0.015, MAE = 0.096 and RMSE = 0.125, training accuracy as 84.9% and test accuracy as 64% which suggested some degree of overfitting. Narrowing down

to the 6 most seasonal features and training several other models yielded drastically improved results. Results from the 4 best performing algorithms are presented below.

1. Logistic Regression: MAE = 11.68, MSE = 267.68, RMSE = 16.36.
2. SVM: MAE = 16.58, MSE =433.33, RMSE = 20.81
3. KNN: MAE = 9.68, MSE = 136, RMSE = 11.66
4. Rand Forest: MAE = 10.86, MSE = 201.38, RMSE = 14.19

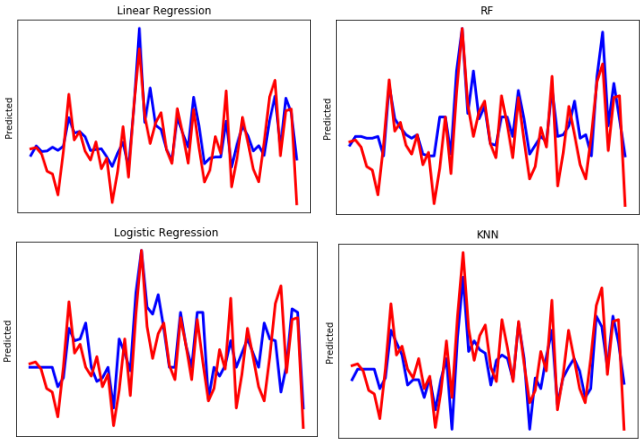


Fig. 6 Performance Comparison of Machine Learning Algorithms

The graphs above illustrate the predicted plot in blue and actual plot in red. When the number of features were reduced to 6 by eliminating the 5 least reliable datasets, the Training accuracy fell to 82% and the test accuracy increased to 69%. Reducing the number of features in this case reduced overfitting and yielded an increase in the test accuracy. The best performing algorithm, logistic regression using the 6 most reliable augmented features, was able to outperform the 3 year and 5 year S-ARIMA time series predictors, with an MSE (Mean Squared Error) of only 0.22. This far outperforms the industry standard.

View/try our code on Github [15].

V. Conclusion

With the increasing availability of energy data and the emergence of simple yet powerful machine learning algorithms in recent years, energy commodity forecasting has become easier and more accurate than ever before. Our findings suggest that state of the art analysis in this space can be conducted using widely available open source feature discovery techniques and machine learning packages.

A difficult aspect of energy forecasting, and a fundamental limitation of this technique, is that there are massively influential factors that simply cannot be predicted; the political/economic climate, corporate decisions, weather patterns, OPEC and international policy changes all drive the price of electricity on a daily basis.

One solution we would like to explore 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 (in Economics or Politics, for instance). Using the time series data augmentation method described above, we hope to develop a dynamic model capable of changing weights and adding or adjusting the influence of features over a particular time bracket in the future, and providing several possible outcomes based on qualitative parameters like the probability of beneficial or adversarial political or economic events occurring at a given time.

Conceivably, this model would utilize online learning from databases updated in real time providing constant feedback into the pipeline, and would be able to translate real world economic changes to a corresponding change in the feature set and weights used in the model. This would allow the user to consider different scenarios by changing the contributing factors of the prediction. Logical next steps in this direction would be to develop a consistent method for quantifying perturbations, and to investigate how to incorporate and translate domain knowledge into changes in the prediction algorithm's parameters.

References

[1] DiChristopher, T. (2018). US energy industry is swimming in data they don't use. [online] CNBC. Available at: https://www.cnbc.com/2015/03/05/us-energy-industry-collects-a-lot-of-operational-data-but-doesnt-use-it.html/.

[2] "Kaggle: Your Home for Data Science", Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/.

[3] "data.ca.gov", Data.ca.gov, 2018. [Online]. Available: https://data.ca.gov/.

[4] "Python Data Analysis Library — pandas: Python Data Analysis Library", Pandas.pydata.org, 2018. [Online]. Available: https://pandas.pydata.org/.

[5] J. Brownlee, "Autoregression Models for Time Series Forecasting With Python - Machine Learning Mastery", Machine Learning Mastery, 2018. [Online]. Available: https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/.

[6] "StatsModels: Statistics in Python — statsmodels 0.8.0 documentation", Statsmodels.org, 2018. [Online]. Available: http://www.statsmodels.org/stable/index.html.

[7] Alvarez, F.M., et al., Energy Time Series Forecasting Based on Pattern Sequence Similarity. IEEE Transactions on Knowledge and Data Engineering, 2011. 23(8): p. 1230-1243.

[8] Hiroyuki, M. and A. Akira. Data mining of electricity price forecasting with regression tree and normalized radial basis function network. in 2007 IEEE International Conference on Systems, Man and Cybernetics. 2007.

[9] Neupane, B., et al. Artificial Neural Network-based electricity price forecasting for smart grid deployment. in 2012 International Conference on Computer Systems and Industrial Informatics. 2012.

[10] González, C., J. Mira-McWilliams, and I. Juárez, Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests. IET Generation, Transmission & Distribution, 2015. 9(11): p. 1120-1128.

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

[12] Zareipour, H., et al., Classification of Future Electricity Market Prices. IEEE Transactions on Power Systems, 2011. 26(1): p. 165-173.

[13] Sadeghi-Mobarakeh, A., et al. Data mining based on random forest model to predict the California ISO day-ahead market prices. in 2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). 2017.

[14] Weron, R., Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, 2014. 30(4): p. 1030-1081.

[15] https://github.com/afbholly72/Energy_Price_Predictor