Long-Term Electricity Price Forecast Using Machine Learning Techniques

Ashkan Yousefi
UC Berkeley
Berkeley,California
ashkan.yousefi@berkeley.edu

Omid Ameri Sianaki
Victoria University
Sydney, Australia
Omid.AmeriSianaki@vu.edu.au

***Abstract*—Predicting the performance of energy commodities has long been a global priority for researchers and investors in the Energy sector. Large green field and brown field 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 accurate insights for the investors. The new approach was proposed in this research to provide data analytics backed analysis for the performance of energy related commodities using innovative 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 energy market. Data points from 2001 to 2017 were collected and 78 data points are considered for analyses to select the highly-correlated features which could potentially affect the electricity price in the medium to long term. An economic case study is undertaken to understand the correlation of the features, and to avoid multicollinearity. In the next step, the selected features are applied into an S-ARIMA time series prediction algorithm. In addition, several feature-based machine learning algorithms are applied to the data and the results analysed and compared to find the effective forcasting approach. The findings demonstrated promising results for three years future price prediction horizon. Further studies are required to get more accurate electricity results beyond three years horizon.**

Keywords— Machine Learning Algorithms, Clustering, Random Forests, Time Series Analysis, Feature Correlation Analysis, Energy Price Prediction

I. INTRODUCTION

Prior to 2008, investments in large power-plant and energy infrastructure projects were largely led by big banks in the USA. Since the 2008 market crash, however, investments in these assets reduced significantly, 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 the sector started to attract new investors for the capital intensive energy assets. This trend created a gap in the market between the significant demand for innovative energy projects with great potential for return and investors. The main reason for disconnected investment scene exists, both domestically and abroad, mostly as a result of lack of information and uncertainty about the projected financial performance and the main driving factor for financial return is the price of energy in the long term.

The present research aims to get advantage of new powerful data analytics tools and techniques to produce a reliable prediction for the performance of energy particularly the monthly retail price of electricity 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 the energy investment space. In addition to applications in institutional investments, the proposed methodology could aid everyday economists and researchers in making their forecasts and create a noticeable impact on policy makers in energy industry to determine the penetration level for demand response programs and other emerging sustainable energy sources.

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. However, the energy projects and the required data sets are not available comparing with the lage online and ofline collection of data in the energy industry. The problem is that there exists a massive gap between the analysis currently conducted and relied upon in the field, and the analysis that could potentially be done given the abundance of online and ofline data available in the energy industry. In addition, emergence of simple and powerful machine learning techniques in the last few years

initiated a trend in the traditional energy industry to get benefited from the massive pool of data to enhance the offered services and bring new opportunities for investors in the energy sector. The present research 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 lead to data driven decision making for the investors and to help them to achieve more clarity in the next three to five years. Previous researches related to the energy forecast focused on the load forecast. The reason is that the load forecast traditionally drives the investments for the energy infrastructure expantions and also the requirements for building new power plants was driven from the load forecast. In addition, the load forecast indirectly can affect the price of energy in the specific market. Some of the previous researches and the applied methods for the forecast is summarized in the following:

Rahul [2] presented a novel method for long term load forecasting with hourly resolution. The model is fundamentally centered on Recurrent Neural Network consisting of Long-Short-Term-Memory (LSTM-RNN) cells. The proposed model is found to be highly accurate with a Mean Absolute Percentage Error (MAPE) of 6.54 within a confidence interval of 2.25% which is favorable for offline training to forecast electricity load for a period of five years.

Hossein investigated [3] the problem of long-term load forecasting for the case study of New England Network using several commonly used machine learning methods such as feedforward artificial neural network, support vector machine, recurrent neural network, generalized regression neural network, k-nearest neighbors, and Gaussian Process Regression. The results of these methods are compared with mean absolute percentage error (MAPE).

Arild presents a framework for price forecasting in hydro-thermal power systems. The presented framework consists of a long-term strategic and a short-term operational model. This research, facilitate more detailed fundamental market modeling to enable realistic multi-market price forecasting. In addition, some of the technical constrained on the price such as cable ramping and reserve capacity are also considered for the price forecast [4].

The impact of load forecast on electricity pricing is considered in the research done by Baifu. In this research, a load forecasting model considering the Costing Correlated Factor (CCF) with deep Long Short-term Memory (LSTM). Also, this paper uses an adaptive Moment Estimation algorithm for network training and the type of neuron is Rectified Linear Unit (ReLU). Baifu concluded that LSTM with CCF can reduce energy cost with acceptable accuracy level [5].

LianLian presented an efficient method for the day-ahead electricity price forecasting (EPF) based on a long-short term memory (LSTM) recurrent neural network model. The applied method is capable of learning features and long term dependencies of the historical information on the current predictions for sequential data. The proposed method is successfully applied for Australian market at Victoria (VIC) region and Singapore market [6].

Hongqiao proposed a robust predictive model construction and probabilistic forecasting with low-resolution data. In this research, the combination of high and low resolution data lead to more effective probabilistic forecast and improve the outcomes [7].

II. Analysis and Methodology

A. Data Collection and Cleaning

Data from the US Energy Information Administration's OpenData database [8] (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 data points 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 [9] 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 [10] model was trained via the autoregression model in the Statsmodels Python Package 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 the following section.

III. CASE STUDY

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 California. Inconsistencies were found in almost every data table obtained from merging data from various sources. Finally, reliable energy data set is prepared after the comprehensive data cleaning in the Energy Information Administration's OpenData database.

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. Numbers indicate correlation of feature with Average Monthly Retail Electricity Price:

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

The correlation of unrelated factors that were highly correlated with the average price of electricity were surprising. The interesting findings was related to the high correlation of the consumption and production of renewable energy features to the retail price of electricity including solar, geothermal, wind, and biofuel and basically overall renewable generation were heavily correlated with the retail price of electricity. Our findings seems to agree that there exists a 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.

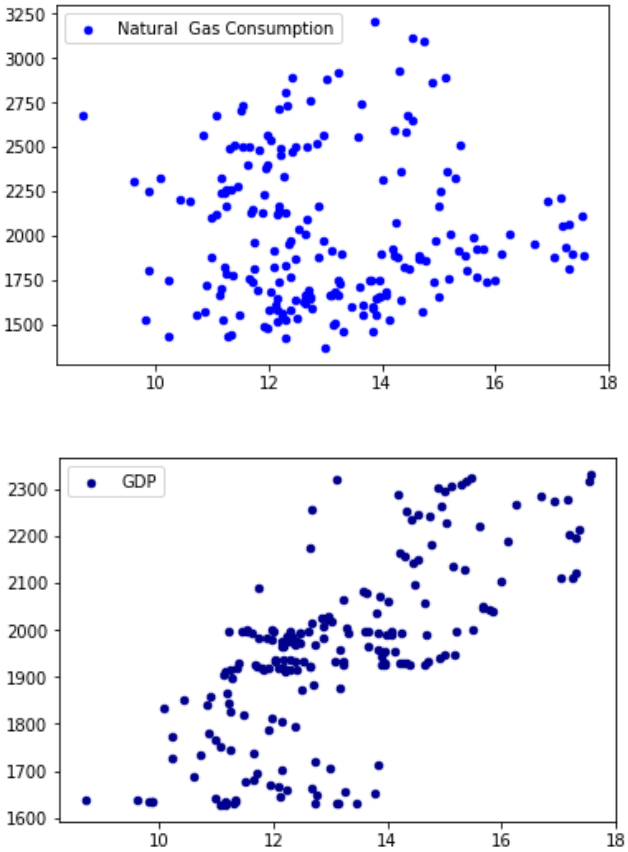

Fig 2: Natural Gas Consumption and GDP vs. Electricity Price

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.

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.

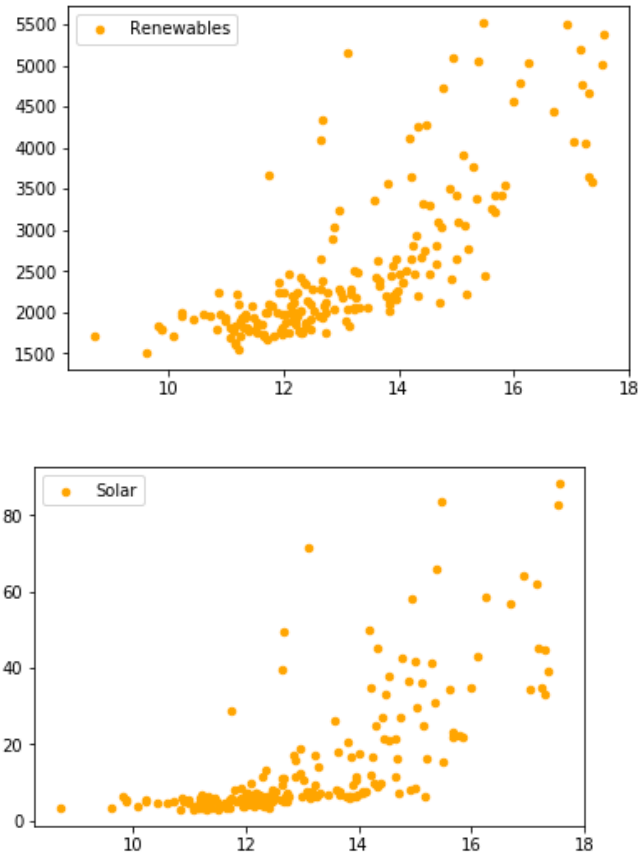

Fig 3: Renewables and Solar vs. Electricity Price

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.

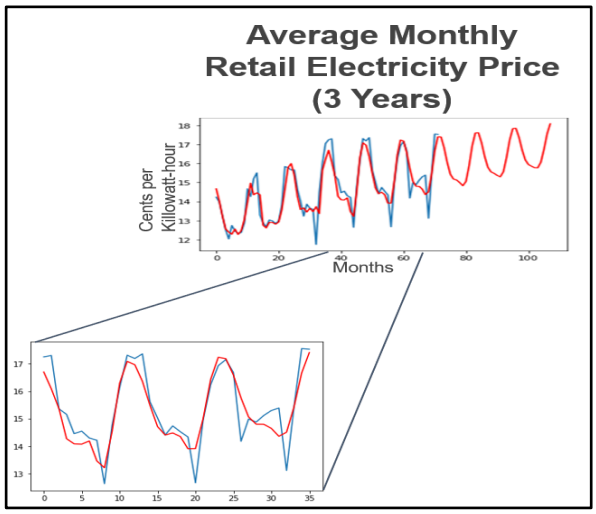

Fig. 4 Performance of 3-Year ARIMA Predictor

C. 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

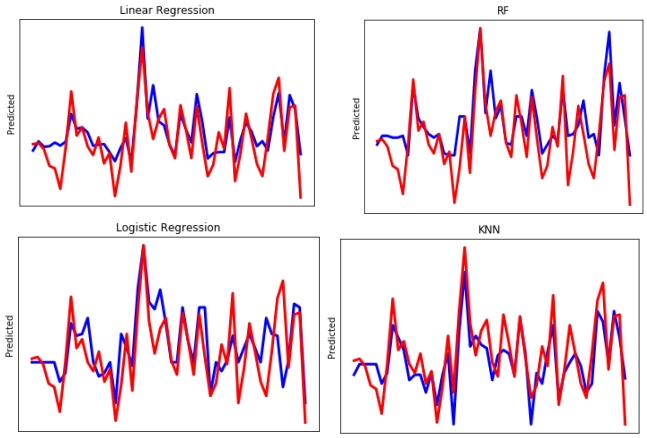

Fig. 5 Performance Comparison of Machine Learning Algorithms

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 less 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 in better test accuracy. The best performing algorithm, logistic regression using the 6 most reliable 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. The achived results outperforms the reviewed literature and the industry standard.

IV. Conclusion

The availability of energy data and the emergence of simple yet powerful machine learning algorithms in recent years, energy price 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 data sources, feature discovery techniques and machine learning. 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, 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] D. C. Maheepala, R. M. N. Nayanajith, M. W. R. P. Somarathna, R. A. A. M. Bandara, and K. T. M. U. Hemapala, "Designing an Energy Monitoring, Analysing and Solution Providing System for Energy Auditing," in 2018 Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2018, pp. 1-5.

[2] R. K. Agrawal, F. Muchahary, and M. M. Tripathi, "Long term load forecasting with hourly predictions based on long-short-term-memory networks," in 2018 IEEE Texas Power and Energy Conference (TPEC), 2018, pp. 1-6.

[3] H. Sangrody, N. Zhou, S. Tutun, B. Khorramdel, M. Motalleb, and M. Sarailoo, "Long term forecasting using machine learning methods," in 2018 IEEE Power and Energy Conference at Illinois (PECI), 2018, pp. 1-5.

[4] A. Helseth, M. Haugen, S. Jaehnert, B. Mo, H. Farahmand, and C. Naversen, "Multi-Market Price Forecasting in Hydro-Thermal Power Systems," in 2018 15th International Conference on the European Energy Market (EEM), 2018, pp. 1-5.

[5] B. Huang et al., "Load Forecasting based on Deep Long Short-term Memory with Consideration of Costing Correlated Factor," in 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), 2018, pp. 496-501.

[6] L. Jiang and G. Hu, "Day-Ahead Price Forecasting for Electricity Market using Long-Short Term Memory Recurrent Neural Network," in 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), 2018, pp. 949-954.

[7] H. Peng, C. Liu, K. Bai, and J. Gu, "Methods of Improving Annual Energy Forecasts With Low-resolution Data," in 2018 IEEE Power & Energy Society General Meeting (PESGM), 2018, pp. 1-5.

[8] (2017). California Government Open Data Portal. Available: https://data.ca.gov/

[9] "Python Data Analysis Library," ed, 2017.

[10] J. Brownlee, "Autoregression Models for Time Series Forecasting With Python - Machine Learning Mastery," ed, 2018.