**Capstone Three Report**

Springboard Data Science Course

*Bence Takacs*

**Problem Statement**

In this project I sought to predict the average price of energy to ultimate consumers in the commercial sector based on data from the U.S. Energy Information Administration. This data includes a wealth of monthly recent and historical energy statistics, including production, consumption, price, and specific sectors like renewable energy and electricity. Successful modeling and prediction of energy prices would allow commercial entities to make electricity price-based decisions, like rationing power in times of higher-than-normal prices or foregoing an energy-expensive venture which doesn’t have large margins. Such information and practices could increase profit margins for businesses.

**Data Wrangling**

The data contained 113 spreadsheets of information. To narrow this down I chose five tables (including the one containing the target feature, price) which I believed could be related to energy price. Table 9.8 contained the average prices of electricity to ultimate customers in several sectors, of which I chose to focus on commercial. The data begin monthly from 1976, although there are several values missing from 1984-1989, so I chose to keep only values from 1990-2024. Running tests for outliers, I found none. The four additional tables of information I chose were table 1.3, primary energy consumption by source, table 3.8a, residential and commercial sectors, table 4.3, consumption by sector, and table C1, population, U.S. GDP, and U.S. gross output, from which I chose only population and real GDP. The first three tables were simply merged with the truncated data from 9.8, but data from table C1 were yearly statistics, and so I performed regressions on the data to fill in monthly values. The data from tables 9.8, 1.3, 3.8a, and 4.3 are endogenous data, they influence and are influenced by the average price, and the data from table C1 are exogenous, they influence but are not influenced by the average price.

**Exploratory Data Analysis**

Because this project deals with time series data, stationarity was an important statistic which was checked by a Kwiatkowski-Philips-Schmidt-Shin (KPSS) test. The endogenous data were stationary after differencing once, but the exogenous data were not stationary after repeated differencing. To investigate the relationship between price and the other endogenous variables I plotted them against each other, an example of which is shown in figure 1.

A graph of different colored dots

AI-generated content may be incorrect.

図 1 Total Renewable Energy Consumption and Total Primary Energy Consumption vs Price

From these relationships, as well as investigation with a heatmap, I chose to keep Total Renewable Energy Consumption, Total Petroleum Consumed by the Commercial Sector, Natural Gas Consumed by the Electric Power Sector, and Natural Gas Consumed by the Transportation Sector, Vehicle Fuel, due to their strong apparent correlations with price. These correlations were verified by a Spearman correlation test, as the data were not normally distributed.

**Pre-Processing**

Pre-processing the data consisted of created a new dataframe of differenced values and splitting the data 65-35 for training and testing. Due to the data being a time series, shuffling the training/test data and scaling were unnecessary.

**Modeling**

To evaluate the data, I chose to use several kinds of time series forecasting models, an AutoRegressive Integrated Moving Average (ARIMA) model, a seasonal variant of this which accommodates exogenous variables (SARIMAX), a Vector AutoRegression (VAR) model, the Facebook Prophet model, and a Holt-Winters exponential smoothing model. We forecast 3 years (36 timesteps) and compare on mean average error (MAE) and root mean square error (RMSE).

*ARIMA*

The Akaike Information Criterion (AIC) optimized order for the stationary price data was (*p, d, q*) = (3, 1, 3). This model cannot capture seasonality, and it forecasts a higher trend.

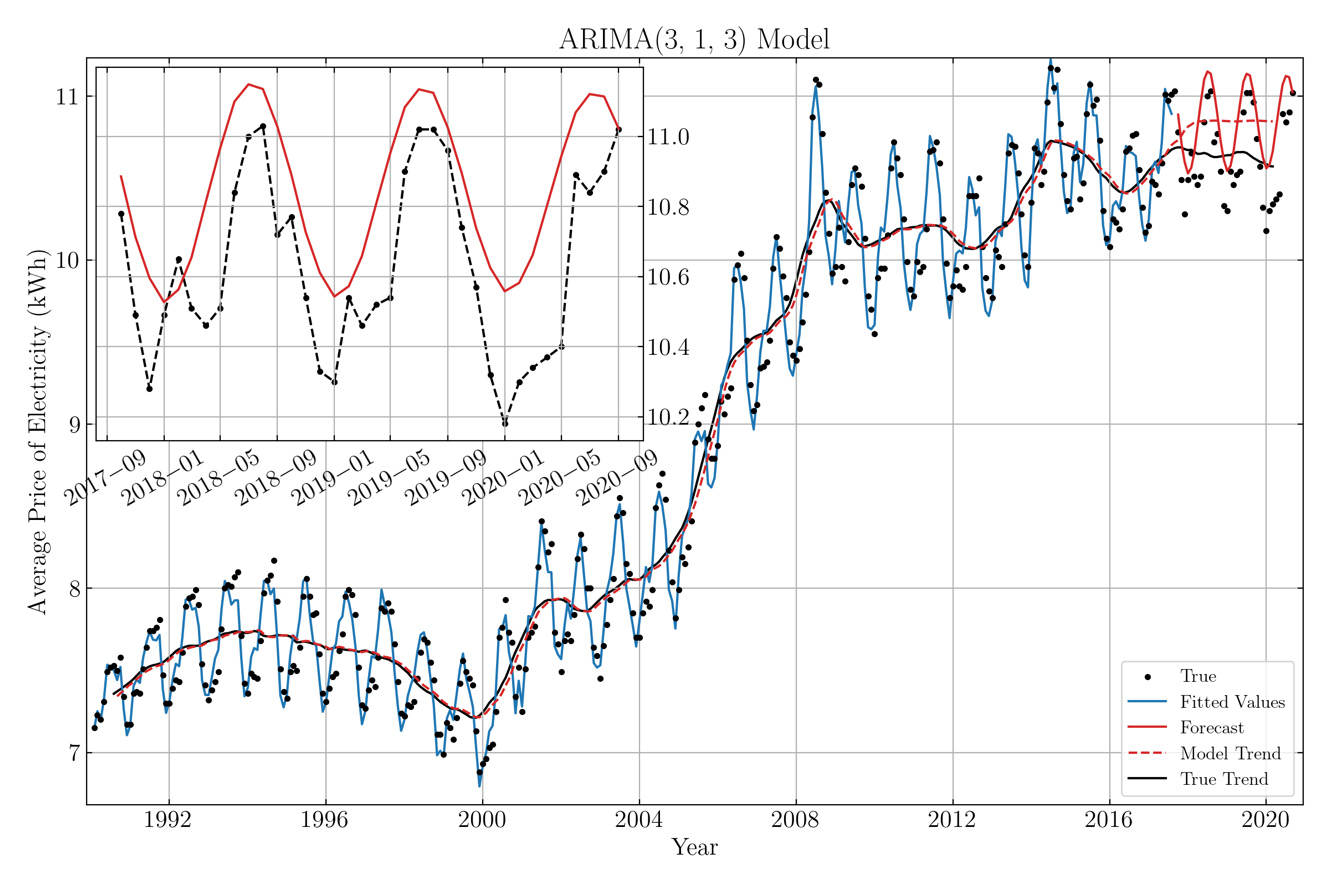


図 2 ARIMA Model Forecast

*SARIMAX*

This model used price data in addition to the exogenous data for fitting. The AIC optimized order and seasonal order is (0, 1, 0) and (1, 0, 1, 12) respectively. This model does very well in forecasting and capturing seasonality, and the trends overlap. It is the best performing model.

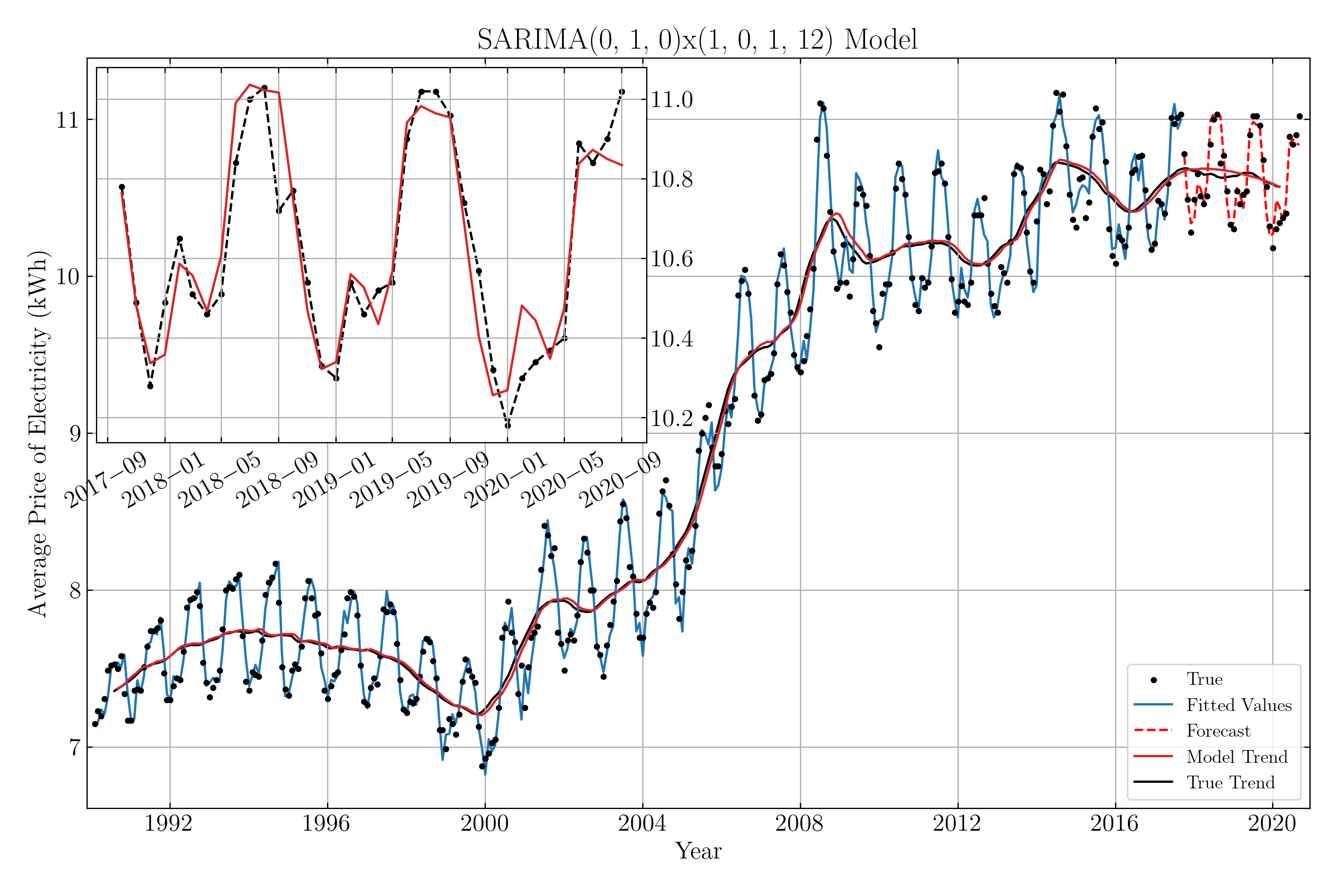
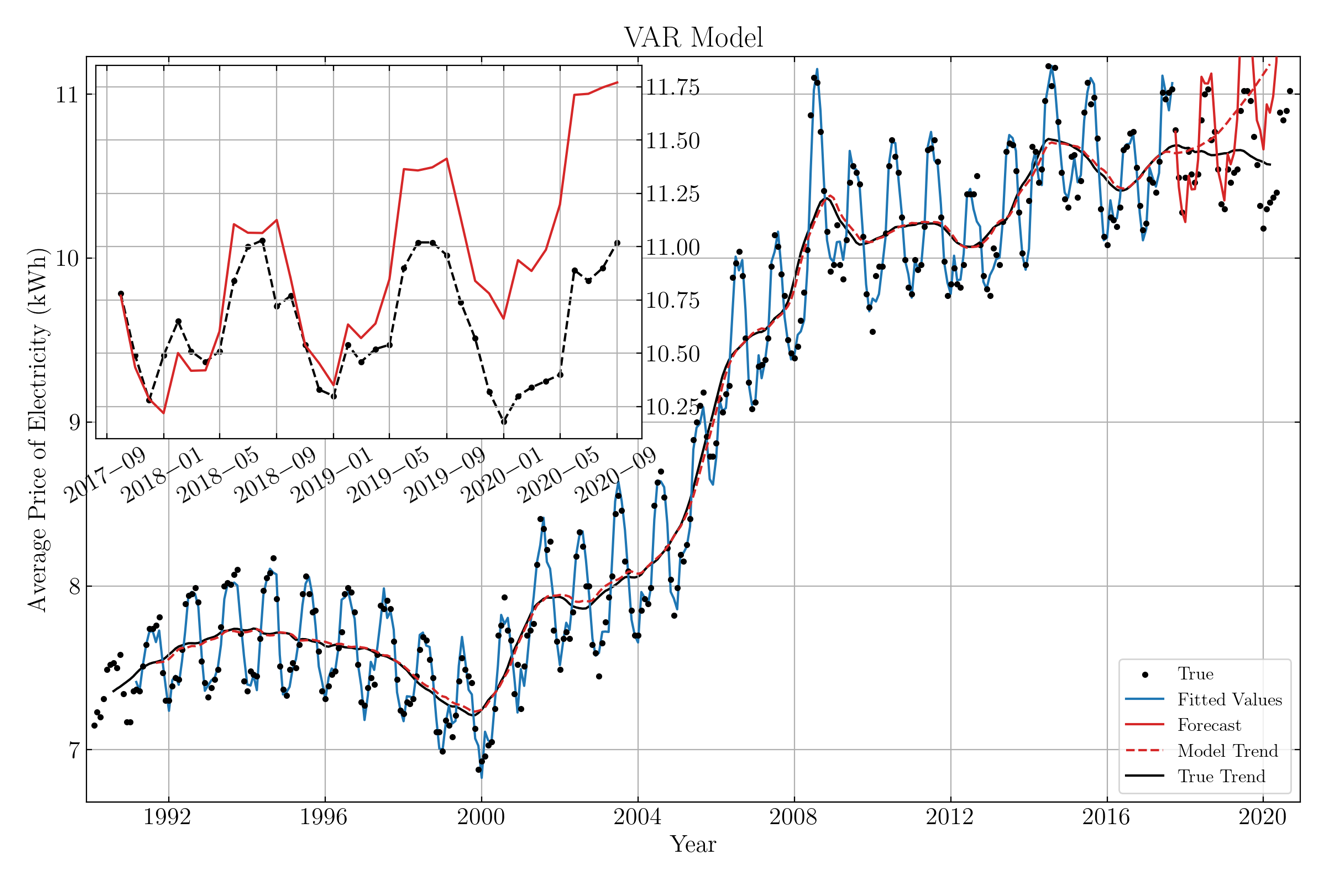


図 3 SARIMAX Model Forecast

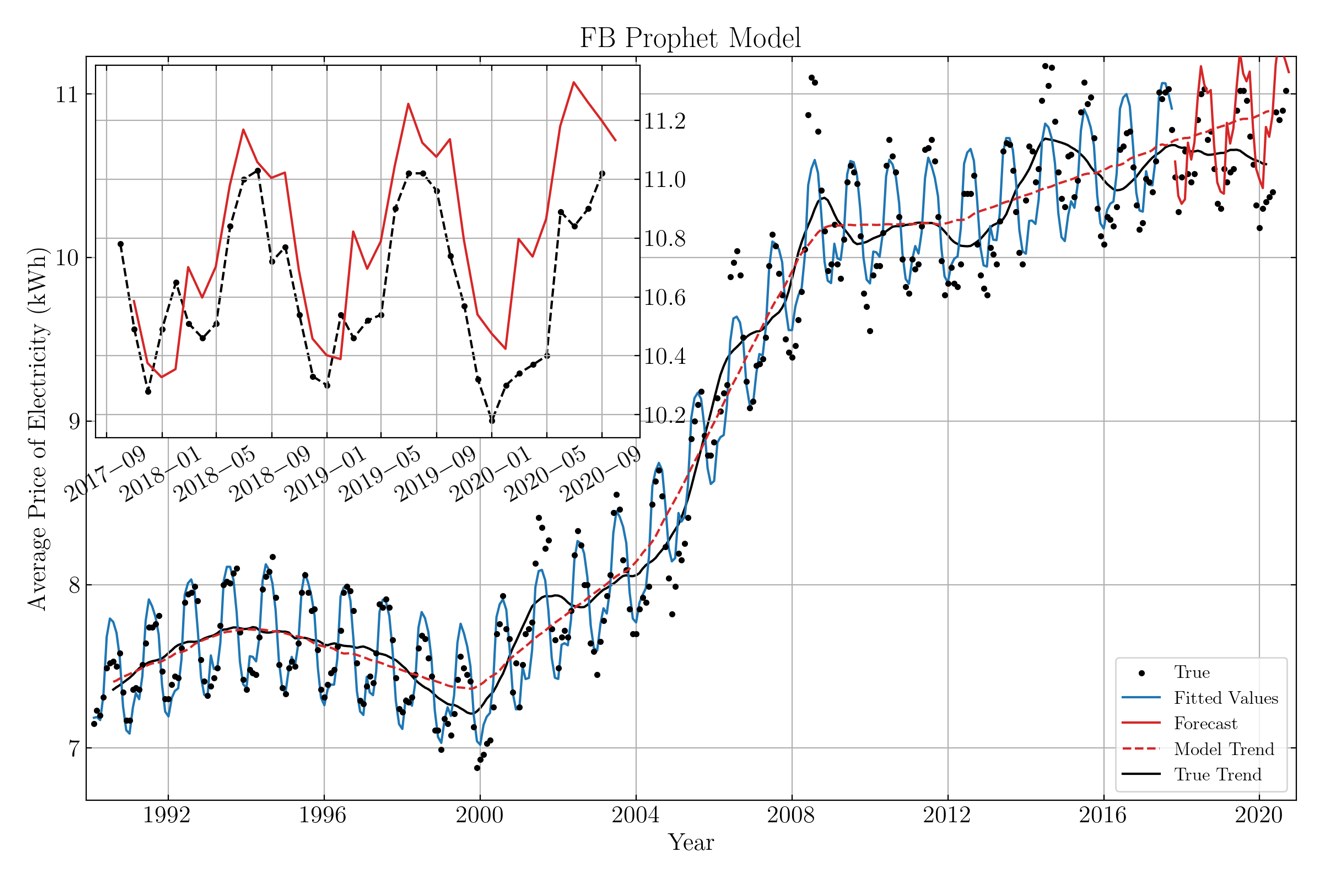
*VAR*

This is a vector autoregressive



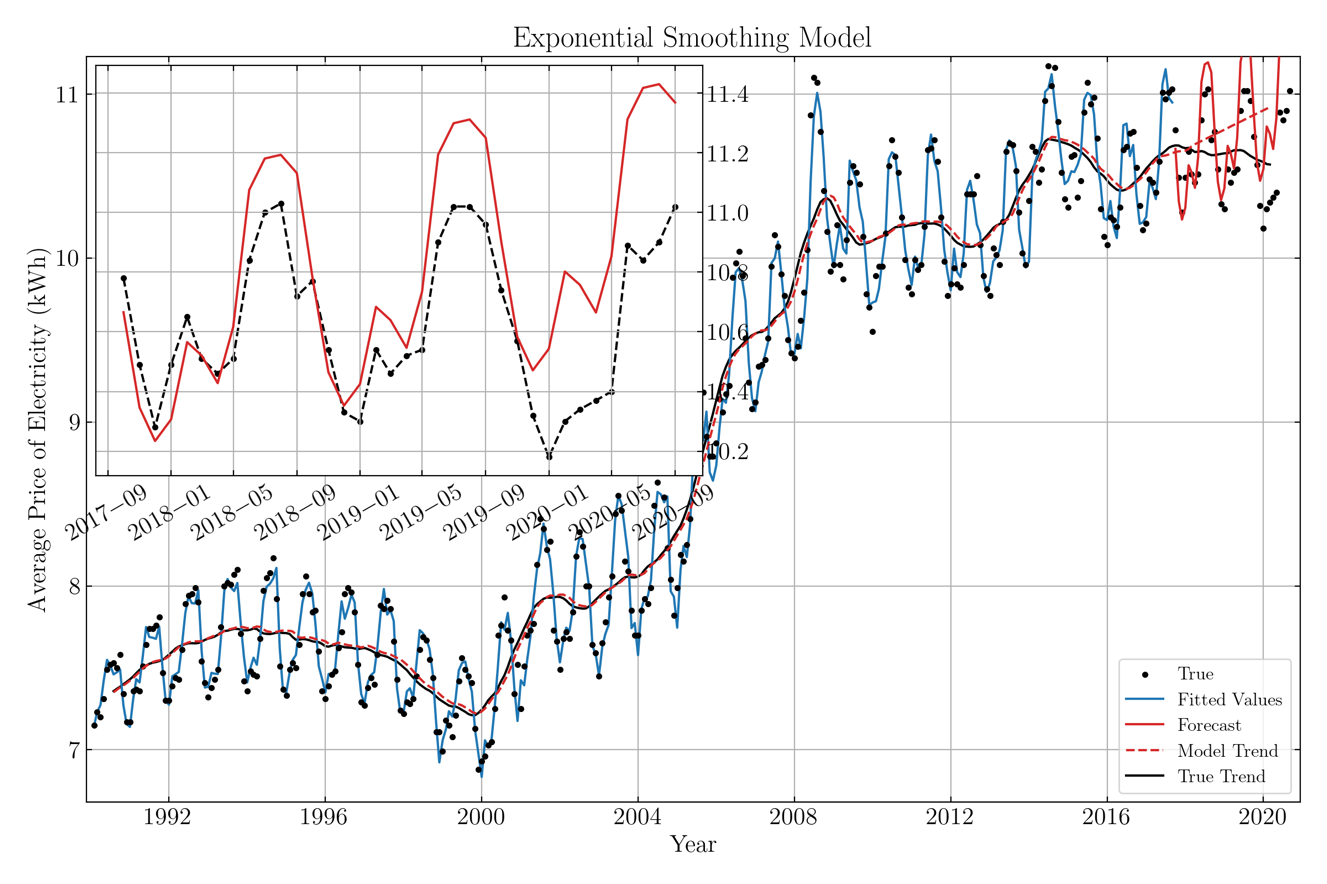
*Facebook Prophet*

The Facebook Prophet model is a thoroughly designed additive time series forecasting model. What makes it unique is its robustness to bad data as well as considering ‘holidays’, or special events which may temporarily impact trends. This is easily seen in the plot, in 2002, 2009, and 2015 the model treats the spikes in energy price as either holidays or outliers and does not try to fit them. This model captures seasonality well but predicts an upward trend.



*Exponential Smoothing*

This model assigns exponentially decreasing weights to past observations, placing importance on recent data. I use the Holt-Winters method as it incorporates seasonality and trend. This model fit the data well and captured the seasonality but trended too high.



**Insert model metrics table here**

Table 1 Model Metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | MAE | RMSE | Information Criterion | Order |
| ARIMA | 0.223 | 0.257 | –321 | (3,1,3) |
| SARIMAX | 0.069 | 0.092 | –556 | (0,1,0)(1,0,1,12) |
| VAR | 0.319 | 0.418 | 6 |  |
| FB Prophet | 0.230 | 0.291 | – | – |
| Exponential Smoothing | 0.219 | 0.271 | –1476 | – |

**Further Research and Recommendations**

Here

**Code and Models**

The finalized code and model metrics can be found in [this GitHub repository](https://github.com/bktakacs/springboard_dsc_capstone3).