

Capstone Three Report

Springboard Data Science Course

Bence Takacs

Problem Statement

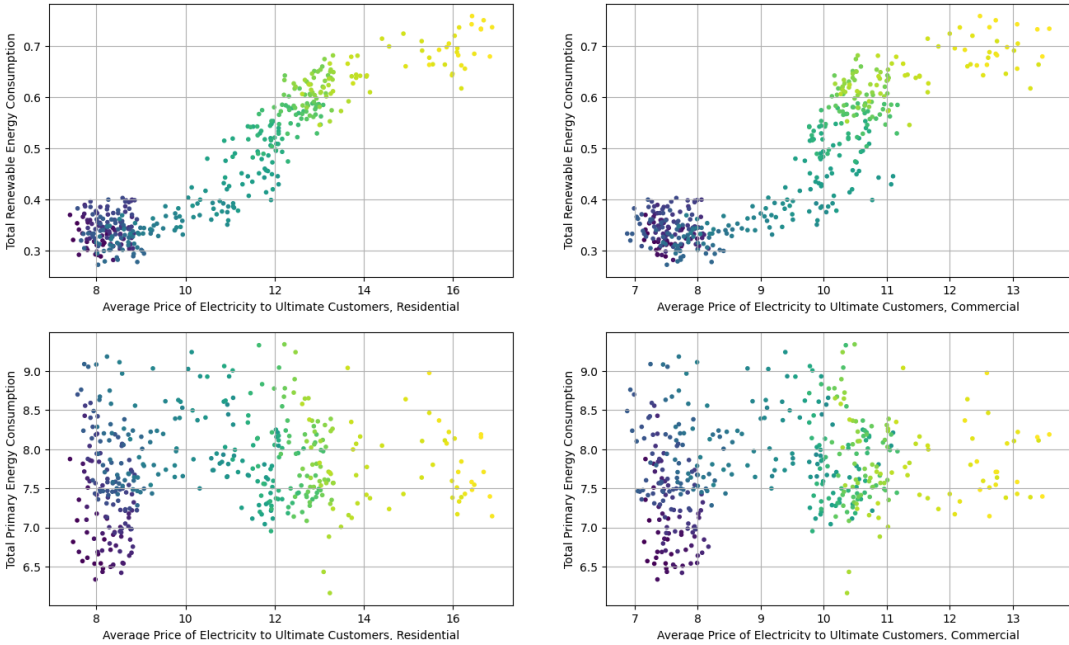
In this project I sought to predict the average price of energy to ultimate consumers in the commercial sector based on data from the U.S. Energy Information Administration. This data includes a wealth of monthly recent and historical energy statistics, including production, consumption, price, and specific 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.

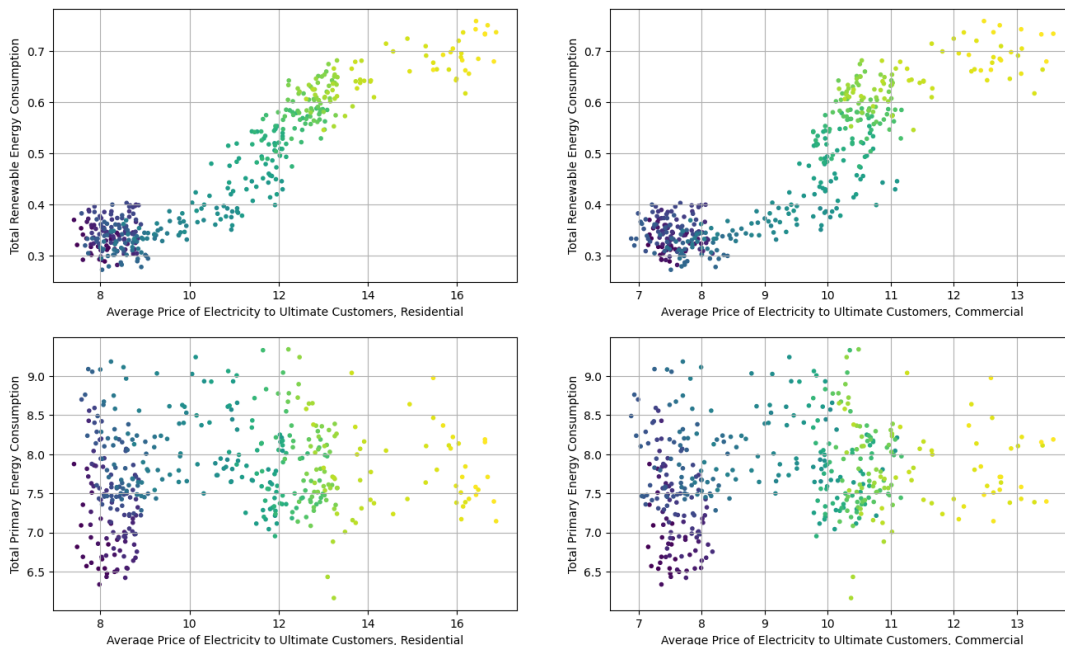
Data Wrangling

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  1. 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.



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

Pre-Processing

Pre-processing the data consisted of created a new dataframe of differenced values and

splitting the data 80–20, respecting chronology, for training and testing, although the forecast only uses about 40% of the test data. This split size was chosen to maximize training data and allow models to make use of more historical data. Because the forecast is short-term, only three years, the remaining 20% of data is plenty to test on. Moreover, in energy and economic forecasting, an 80-20 split is common. Due to the data being a time series, shuffling 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 (HWES) model. We forecast three years (36 timesteps) and compare on mean average error (MAE) and root mean square error (RMSE).

ARIMA

The ARIMA model uses lagged (past) values, integration, and residuals to model a time series. 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 in figure 2.

SARIMAX

This model differs from ARIMA in that it includes seasonality and makes use of exogenous data for fitting and forecasting. 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 in figure 3. It is the best performing model.

VAR

This model can evaluate multiple features at once, and features are modeled using their own and the other past values. Price was not very accurately forecasted in figure 4, but Total Renewable Energy Consumption and Total Petroleum Consumed by the Commercial

Sector were well predicted. Poor prediction may be poor interdependency between endogenous variables, or a lack of exogenous variables.

Facebook Prophet

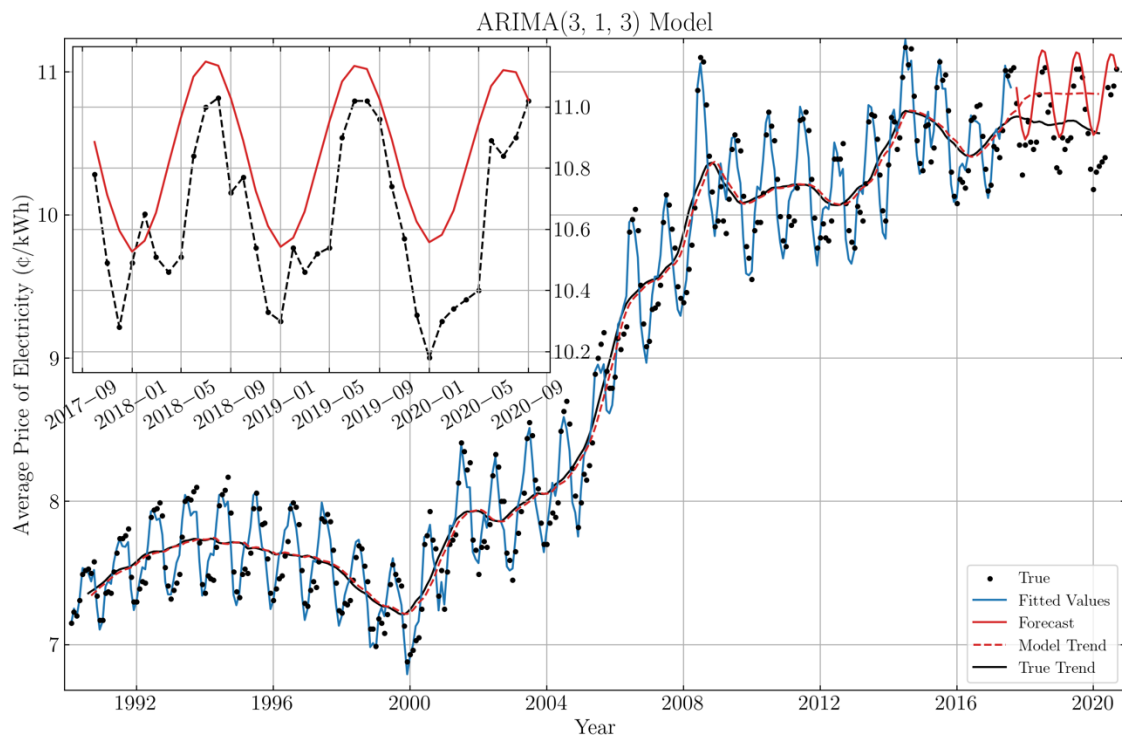
The Facebook Prophet model is a thoroughly designed additive time series forecasting model. What makes it unique is its robustness with bad data as well as considering 'holidays', or special events which may temporarily impact trends. This is clearly seen in figure 5; in 2002, 2009, and 2015 the model treats the spikes in energy price as either holidays or outliers and does not try to fit them, instead fitting to the trend of data before or ahead. This model captures seasonality well but predicts an upward trend.

Exponential Smoothing

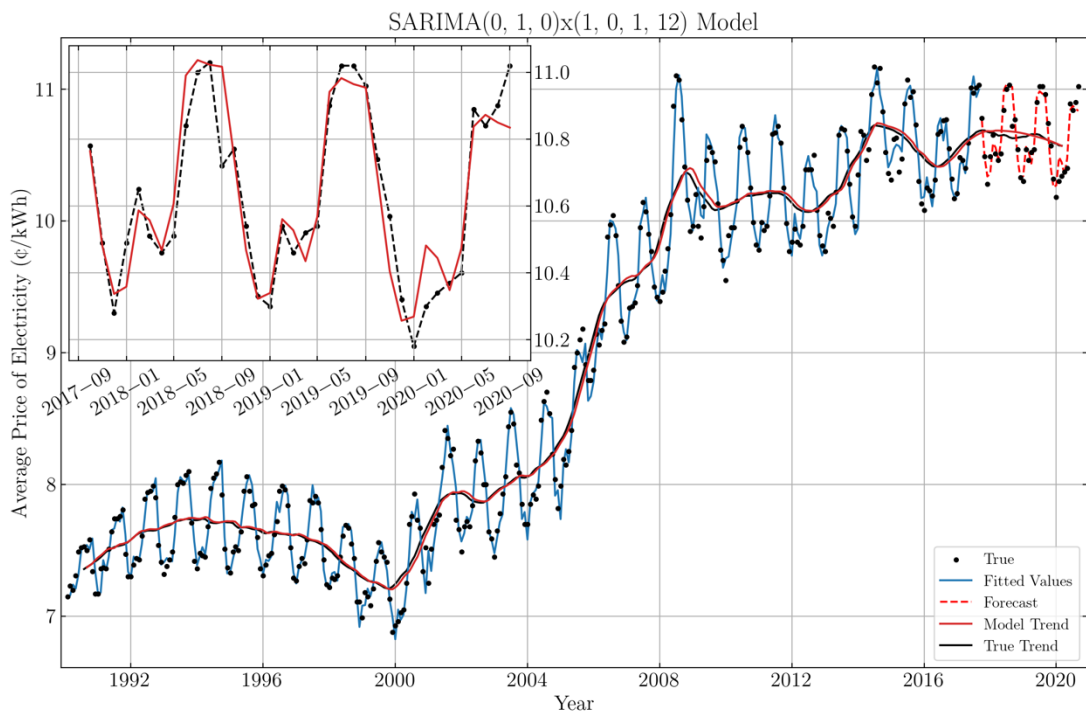
This model assigns exponentially decreasing weights to past observations, placing importance on recent data. I use the Holt-Winters method as it incorporates seasonality and trend. Optimizing model parameters suggested no trend and additive seasonality. This model fit the data well and captured the seasonality but trended too high in figure 6.

One Step Ahead Cross Validation

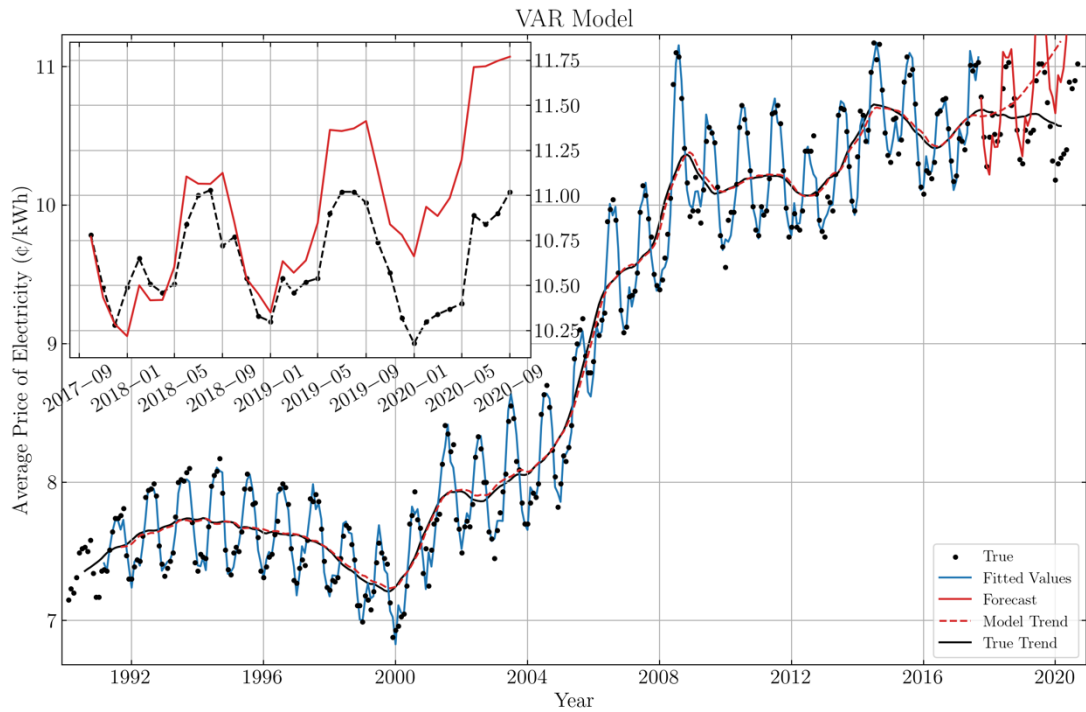
Performing one step ahead cross validation (OSACV) revealed important information about the models. OSACV involves beginning with the train set, predicting the next value, and then including that prediction in the new training set, so that successive iterations have a rolling origin of test data. The MAE and RMSE of OSACV can serve as benchmarks for our models, and can warn of possible over or underfitting, or unoptimized hyperparameters. According to OSACV, the ARIMA and HWES models are underperforming, as their scores do not beat OSACV, but the SARIMAX and VAR models perform well.



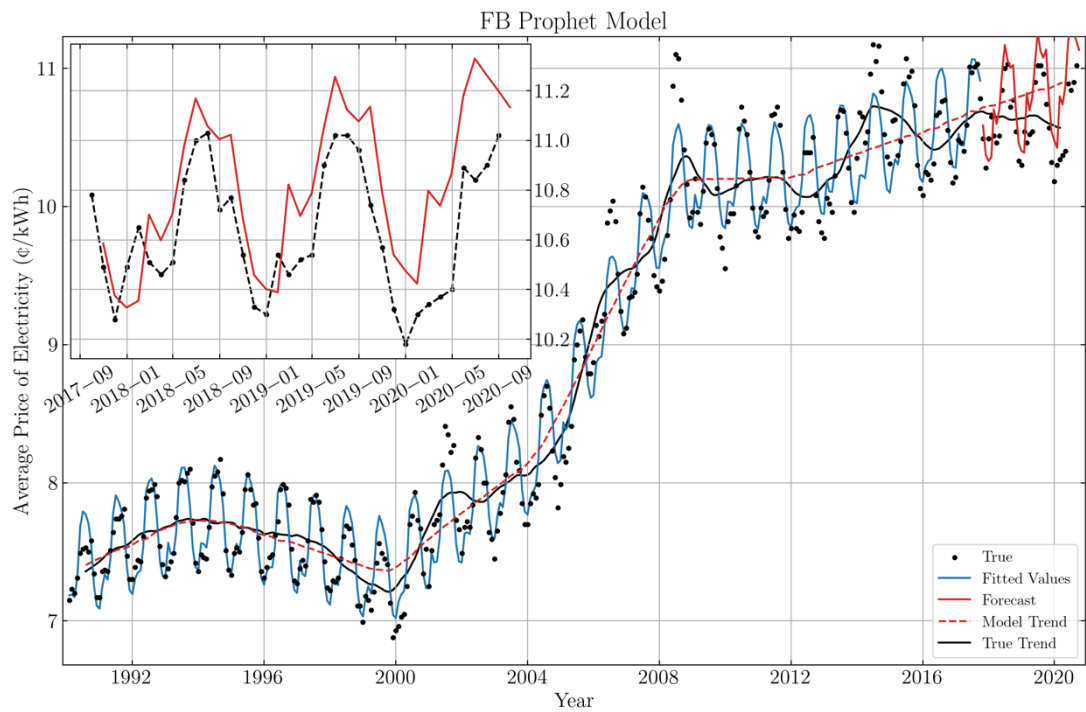
☒ **2 ARIMA Model Forecast**



☒ **3 SARIMAX Model Forecast**



4 VAR Model Forecast



5 Facebook Prophet Model Forecast

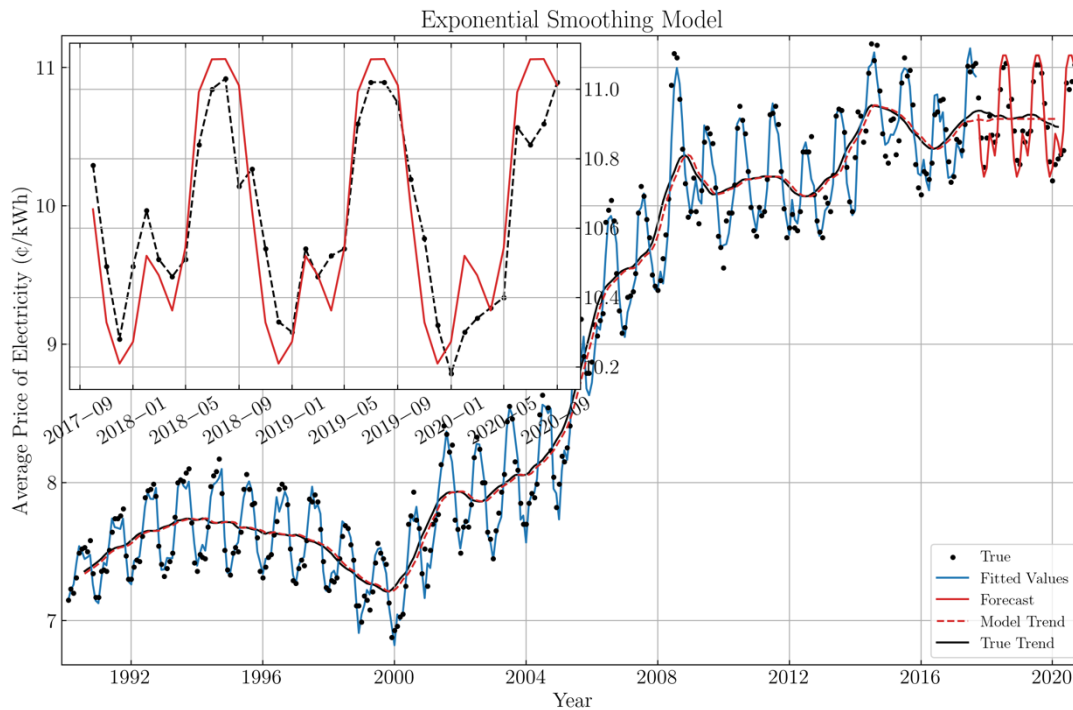


Figure 6 Exponential Smoothing Model Forecast

Table 1 Model Metrics

Model Name	MAE	RMSE	AIC	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	—	—
HWES	0.110	0.133	-1476	—

Further Research and Recommendations

I believe further research in this project should include a large-scale analysis of all the information provided in the dataset, 100+ tables. This would allow for likely stronger relationships to be found in both endogenous and exogenous data, which would improve forecasting accuracy and could also reduce the number of endogenous variables used. I noticed also that the train-test split size, or more accurately the date at which the data were

split, has an impact on prediction accuracy; obviously energy prices are not simple to predict, and so in periods of stability the forecast is accurate, but when prices begin to jump, such as in 2005 or 2021, models have difficulty predicting this. The price of energy is not an isolated system, so more exogenous variables would be useful in making more accurate predictions. The downside to this is that the models need the future values of exogenous features, which would have to be guessed or predicted by another model, further compounding errors in the model and increasing its complexity.

Clients could use these models or findings in three different ways:

1. Companies can forecast energy price to determine when prices will be high and budget accordingly, either by restricting energy usage or waiting until an opportune time to embark on an energy-expensive venture, such as manufacturing.
2. Traders and financial analysts can use the forecast to inform their trading decisions, such as when is most profitable to buy/sell energy futures.
3. Governments and municipal authorities can use the forecast to inform energy policies, infrastructure projects, and subsidies.

Code and Models

The finalized code and model metrics can be found in [this GitHub repository](#).