Predicting Average Monthly Electricity Price of Residential Sector in the United States

Capstone project-2 for Data Science Career Track bootcamp
Benhur Tedros

Summary

Today in our world, there are many electric power industries, which cover the generation, transmission, distribution and sale of electric power to the general public and industry. As world's population, commerce and transportations are expected to grow, the demand for electric power will increase and so does the revenues from the electricity sales. Electricity retailing is the end product of the processes of electric power industry. This industry makes a lot of revenues from electric sales to residential, commercial, and industrial, transportation sectors and others. The other sector refers to activities such as Public Street and highway lighting. The U.S. Energy Information Administration (EIA) collects sales of this electricity and associated revenue, each month, from a statistically chosen sample of electric utilities in the United States. As the change in the electricity price affects people's life, prediction of future electricity price is helpful in visualizing the overall rates for grid electricity and finding out alternative options.

The analysis of the electricity price can help the electric power industry and the government in designing new electricity coverage, improving the existing ones and helping their customers better. State based further prediction would also provide important asset for the pertinent sectors.

Objective of this project

The goal of this capstone project is to predict the rates of electricity price for U.S residential houses.

Data

The dataset for this project was published by U.S. Energy Information Administration and was downloaded from their website. The dataset is comprised of year, month, year_month, data status, Revenue in thousand dollars, Sales in megawatt hours, and price in cents/kwh for residential sector. The data includes the years from January 1990 to August 2017 and can be downloaded from:

https://www.eia.gov/electricity/data/eia861m/index.html

Methods/Approach

I will treat this project as a time series analysis related problem. The following libraries were used for data loading, wrangling, cleaning, data visualization, developing test harness, data analysis, model evaluation etc.

- Pandas
- Numpy
- Matplotlib
- Scikit-learn
- SciPy
- Statsmodels

A. Data Loading and Wrangling

The dataset is stored in MS Excel spreadsheet in CSV format, which was easily loaded into pandas dataframes. It contains 332 observations with 7 data fields with no missing/null values. The data fields include information on year, month, year_month, Data Status, Revenue_dollar, Sales_Mwatt and Price_Centkwh. The year_month were converted to date-time format. For this project's purpose, a subset of year_month and Price_Centkwh were created, and an exploratory data analysis was carried out.

B. Exploratory Data Analysis

Exploratory data analysis for this dataset is useful in determining the trend of the target variable across the given time. The electricity price was plotted against time [1990-2017] to display the overall trend and variance in seasonality (fig.1). It is also good to visualize the fluctuation of the electricity price of each month across the given years (fig.2). The distribution pattern of the dataset was explored using density and Q-Q plots to figure out if the dataset needs some transformation.

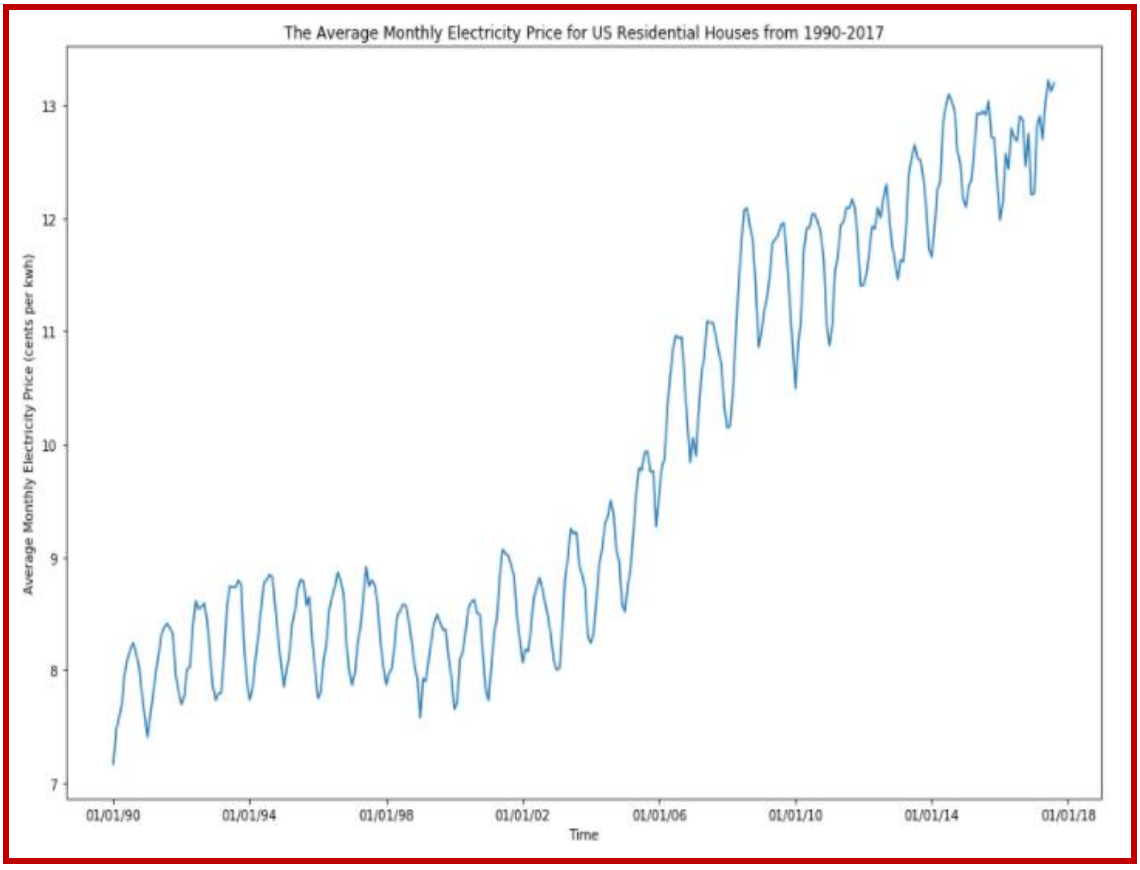

Fig.1 The average electricity price plotted against time

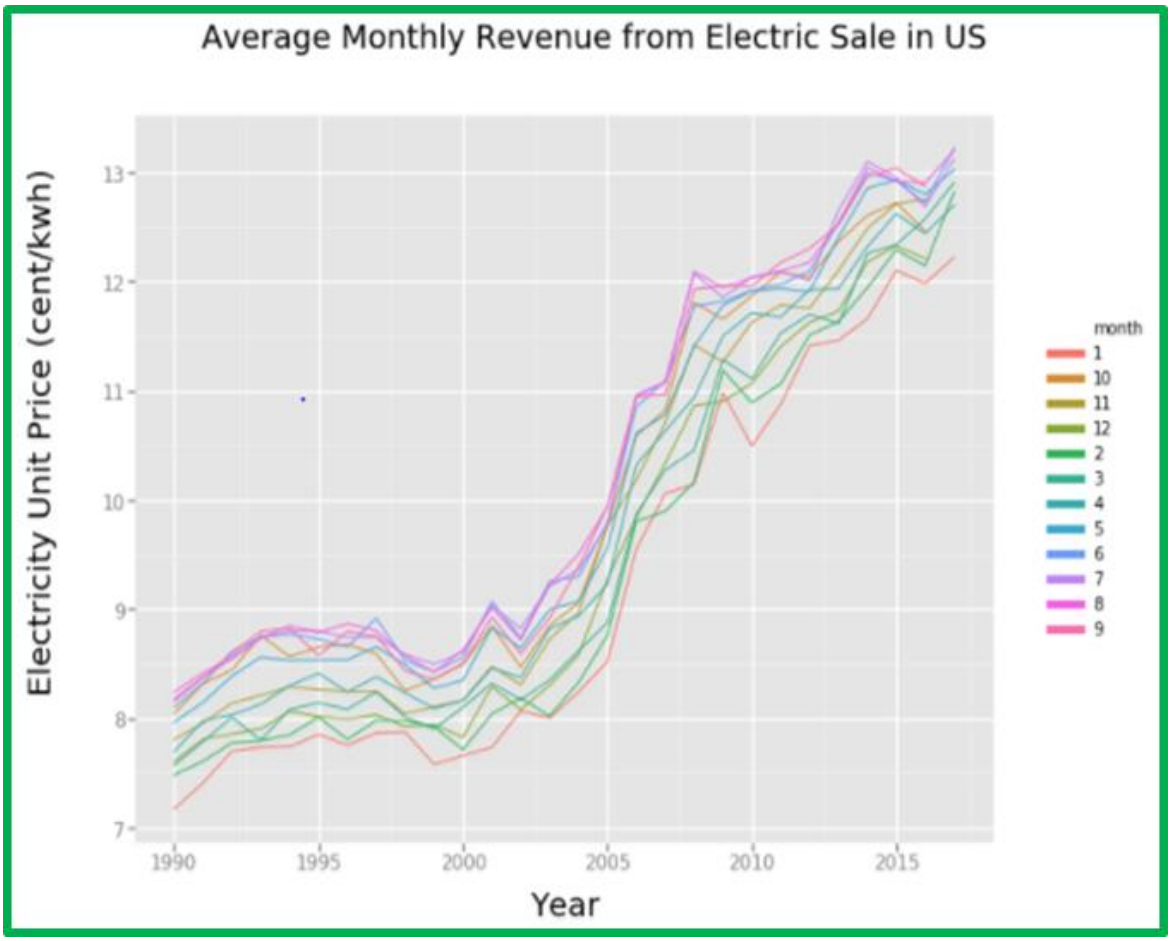

Fig.2 The average monthly electricity price between 1990-2017 across each month

C. Stationarity, Trend and Seasonality in the dataset

When we visualize our time series dataset, it is recommended to look out for few things. The main areas that should be observed are the seasonality, trend and noise in the dataset. Are we seeing any clear periodic pattern in the data; does the data show a consistent trend either upward or downward; is the data have any outlier points which are not consistent with the rest of the data. The rolling statistics and Dickey-Fuller test was carried out to check for the stationarity. As the time series dataset was found to be non-stationary, it was deseasonalized by differencing the electricity values. Once the differenced values were proved to be stationary, the model fitting and prediction was done.

Model Fitting

Time Series algorithms, ARIMA models, was utilized for this project. The predictors depend on the parameters (p, d, q) of the ARIMA model. It is necessary to figure out the values of p and q, for the AR and MA models [ARIMA models] respectively. This was done by plotting ACF and PACF functions. The d value refers to the number of nonseasonal difference. In this case, d was appeared to be zero as the test statistics shows that there is no need for further differencing. Before conducting model fitting, the dataset was split into training [01/1990-06/2003], testing [07/2003-12/2016] and validating [01/2017-08/2017] dataset. The ARIMA models were trained, and a one-step prediction was made. Then, the actual value from the test dataset was added to the training dataset before carrying out the next iteration to test the models. The model with smaller RMSE was chosen to be tested on validating dataset.

Results

ARIMA models were utilized to achieve the objective of this project. With the identified values of p, d and q parameters, the models were trained, tested and evaluated on the given dataset. The p, d and q values were appeared to be 1,0 and 1 respectively. The RMSE of each model was calculated to identify the best one. The AR, MR and ARIMA resulted to a RMSE of 0.3540, 0.3364 and 0.3519 respectively. The MR model was selected to be applied on the validating dataset. Moreover, it is good to review the residual error of the forecast, as its distribution should ideally be gaussian with a zero mean. The residual error from the MR model forecast (considered as bias) was calculated and added to the predicted values for model optimization/correction. The model fit with the calculated bias value was applied to the validating dataset. The result shows a better prediction with the predicted values falling very close to the observed ones. Based on this forecast on the validating dataset, the final RMSE is predicted to be 0.1653 cents per kwh in a month, which was better than that of training/testing dataset (fig.3).

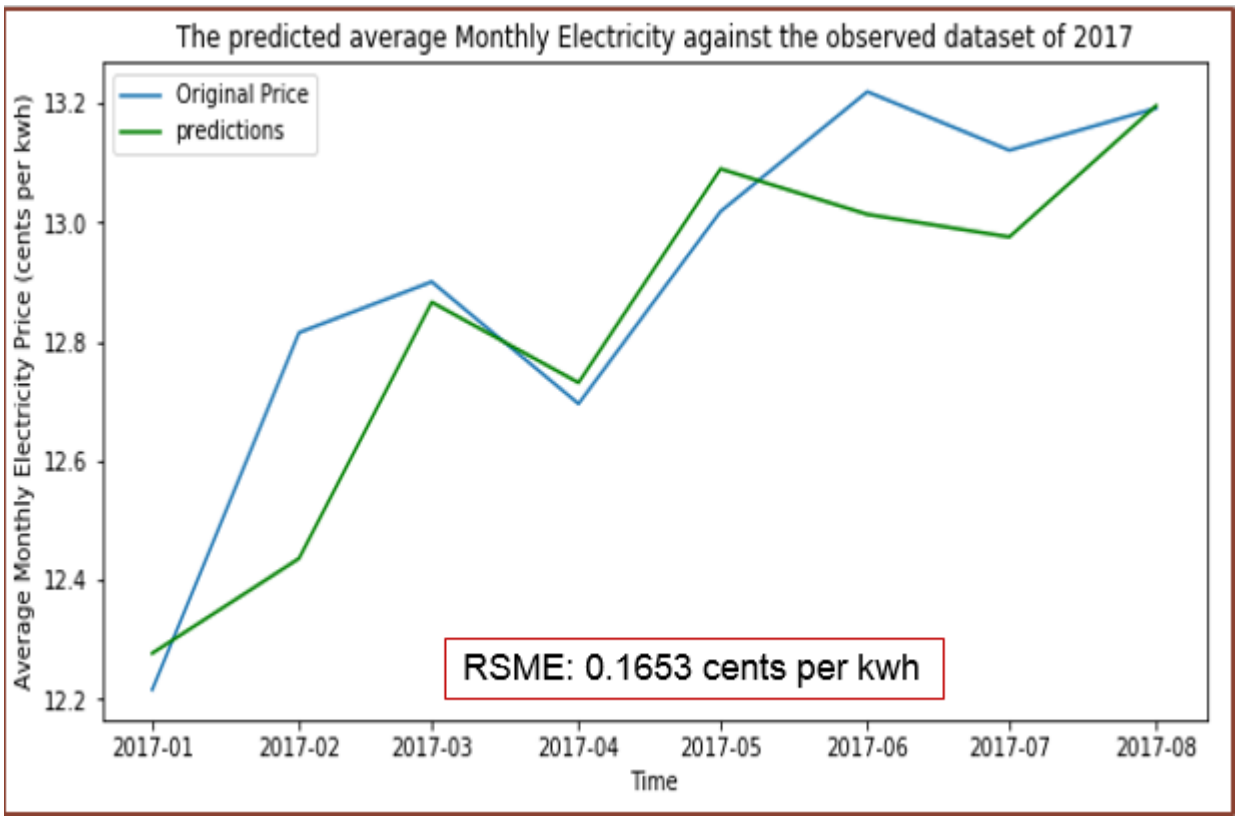

Fig.3 Correlation plot between the predicted and observed electricity price values of the validating dataset

Limitations

The electricity sources might have been different within one state through the given time [1990-2017]. The data does not show whether the given state was using the same type of electricity source throughout that period. Moreover, were there any natural causes/external factors which made the price lower or higher within states? Further, it is good to know how the data was collected and if the same method of data collection was utilized throughout the states.

Further Research

This capstone project tried to predict the average monthly of electricity prices for US residential houses. This research can be broadened by carrying out the prediction for each state independently. From the bi-modal distribution plot, it can be observed that there are two groups of states which bear similar distribution in the electricity prices. As the electric source [natural gas, hydropower, petroleum, coal, solar/wind...etc] differs from state to state, so does the electricity price. Similar research can also be done for commercial, transportation, industrial and other sectors.

Client Recommendations

As the model showed a strong correlation to the dataset, some recommendations can be taken from the carried-out analysis.

- Our client may use the model to forecast the average monthly residential electricity price within the state with an error of 0.1653 cents per kwh.
- It would be helpful if State based prediction has been carried out, as the above analysis is country-based.
- The above analysis assumed that each state had used the same tool and approach to collect their data. It also did not weigh the electric sources' difference from one State to another State. Therefore, the percentage of each electricity source utilized by a state should be weigh and added in the analysis.