Capstone Project 1: Final Report

Title: Prediction of electricity rates (cost/Kwh)

Author: Elizabeth Izarra

Contents

1. General Problem¶

Electricity has a very dynamic market price since it is a commodity that is essential for daily life and non-storable where generation and demand must be continuously balanced. This in turn makes it dependable on the weather conditions.

2. Data Sources¶

- U.S. Energy Information Administration (EIA)
- National Oceanic and Atmospheric Administration NOAA

3. Data Acquisition and Data Wrangling:¶

The data has to be gotten from different web sources. Each source provides APIs which have to be explored to get the required data for the project.

The data acquisition and data wrangling was divided in the following parts:

U.S. Energy Information Administration (EIA)

API Exploration, data acquisition and data wrangling in order to get a single view data set of electricity prices, demand, etc by state per month in a year.

Findings:

- Each variable series has to be independently fetched per State through APIs.

Approach:

- All the data by state corresponding to the same variable was fetched in a loop and concatenated while adding the state information. It was nested in a loop corresponding to all the variables of interest where after fetching each variable, merged them in a single view.
- The rows that do not correspond to one State were deleted.

National Oceanic and Atmospheric Administration (NOAA)

API Exploration, data acquisition and data wrangling of data in order to get a single view data set of temperatures by state per month in a year.

Findings:

- Global Monthly Summary series can be fetched per state through APIs. It provides summary data from each station at a State into a time range. Some States have around 350 stations. Nevertheless, the API maximum limit is 1000 records per fetch.

Approach:

- A couple of loops where nested to fetch data per state per month to cope with the fetching limit. The first loop was used to get the data from all the stations in a State in a

month. It was then grouped by date and variables to get the mean of all the stations in the state during that month. The group was unstacked to be able to concatenate with the information of the following month while adding the corresponding State information. This was nested in a per-state loop.

Merging EIA and NOAA Datasets of one year data

Findings:

- The columns ('date' and 'iso3166'/'State') to be used to merge in a single view both data sets had different formats.

Approach:

- The ISO3166 acronyms for US states with its correspondent states where searched and placed in a csv file.
- On the NOAA temperature data set, a column with the corresponding ISO3166 codes was added, as well as, 'date' was formatted to "YYYY-MM-DD'.
- On the EIA data set, the date was formatted to "YYYY-MM-DD''
- Cleaning functions were defined for both EIA and NOAA data sets.
- Merging of both data sets was done to get a single view of all the data of interest.

Retrieving Data from 2001 to 2018

Findings:

- NOAA server disconnect after a certain time.
- EIA does not have data available through API previous to 2001

Approach:

- Getting the data Year by Year (From 2001 to 2018):
- The three previous steps (1.- EIA date set acquisition per month by state in a year, 2.- NOAA data set Acquisition per month by state in a year, and 3.- Merging of EIA and NOAA data sets) were nested in a loop to get 18 years of data and save them in individuals .csv files per year.
- Getting all years of data in a single view:

All the yearly .csv file were appended to get all the data in a single view and saved into a file.
Files generated:
eia_YYYY_YYYY.csv - one file per EIA year fetched
noaa_YYYY_YYY.csv - one file per NOAA year fetched
data_all_YYYY_YYYY.csv - one file per year fetched with EIA and NOAA data merged
AllData_1.csv - File with all the data_all_YYYY_YYYY.csv files combined

Data Processing

5.1. Checking for missing information:
Findings:

- All States were missing "number of accounts' per month for years early to 2008. In addition the state of Alaska also missed the 'monthly customer accounts' in the year 2016.

Approach:

- A .csv file was found available in EIA website with the mean annual "number of accounts" for each state. This file was downloaded and used as a reference for later interpolation of the "number of accounts", if needed.

5.2. Checking for apparent wrong data:

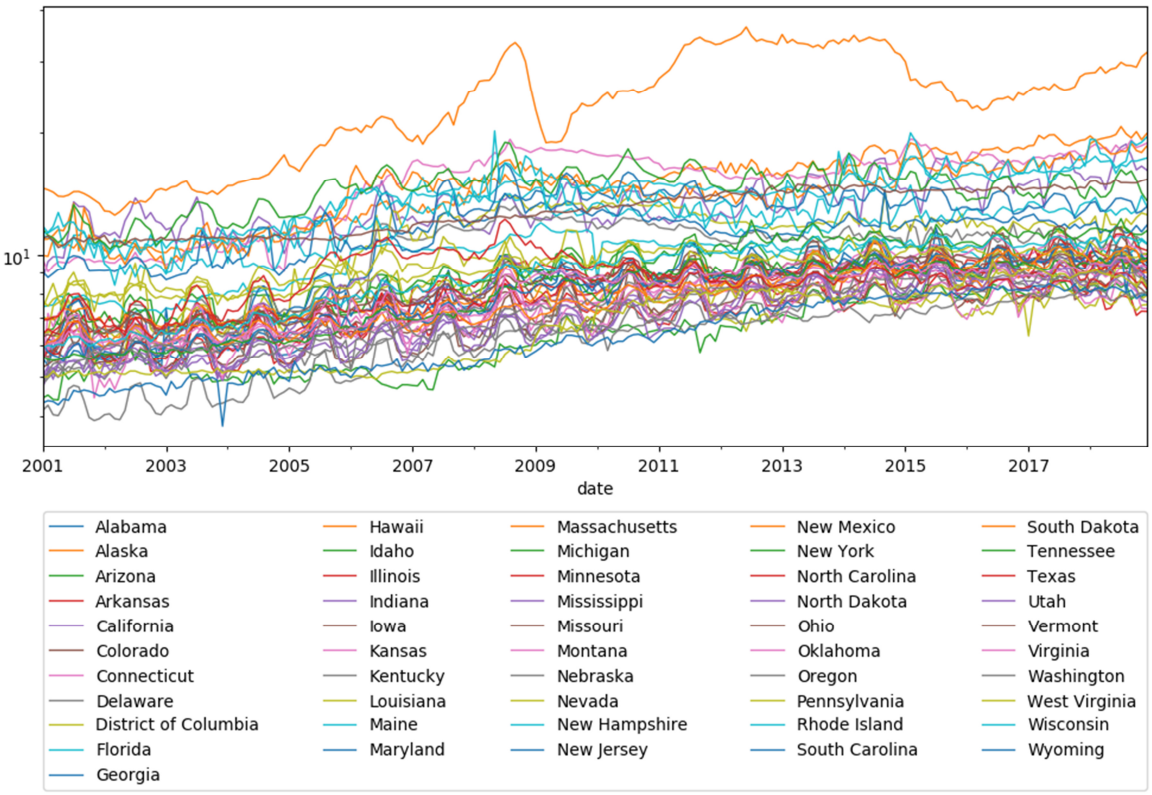
Findings:

- All variables were plotted per State. Graphs show an apparent regular pattern along states with the exception of 'Net generation' where 'District of Columbia' presents some negative values. Retail Electricity Price of "District of Columbia" presents an irregular pattern. (See Figure 1 as an example)

Approach:

- Further analysis will be done before deciding if dropping 'District of Columbia' Data

Figure 1. Average retail of Electricity (cents per kilowatt-hour) – per State



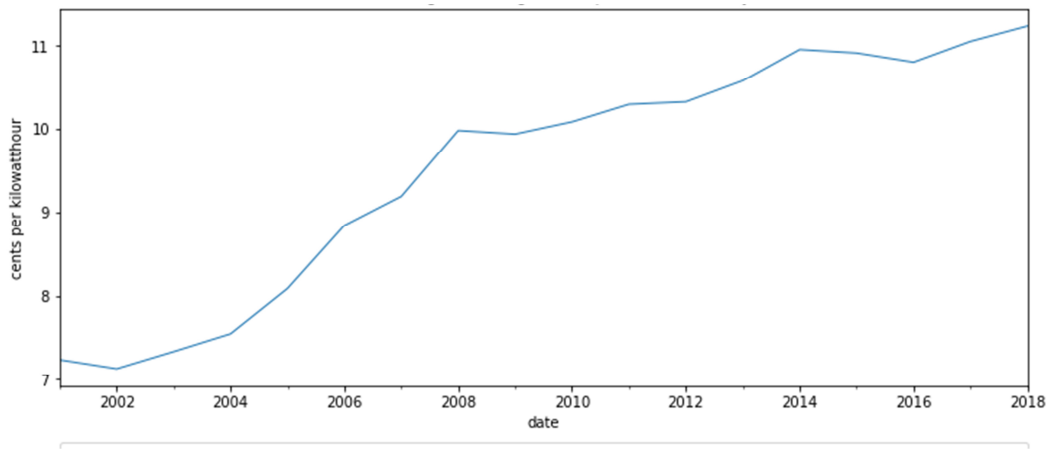
4. Data Visualization¶

Different data visualizations were done in order to identify possible trends, seasonality and dependencies.

Behavior of electricity prices along the years¶

It has been a steady increase in annual aggregated average of electricity price in USA over the last 18 years.

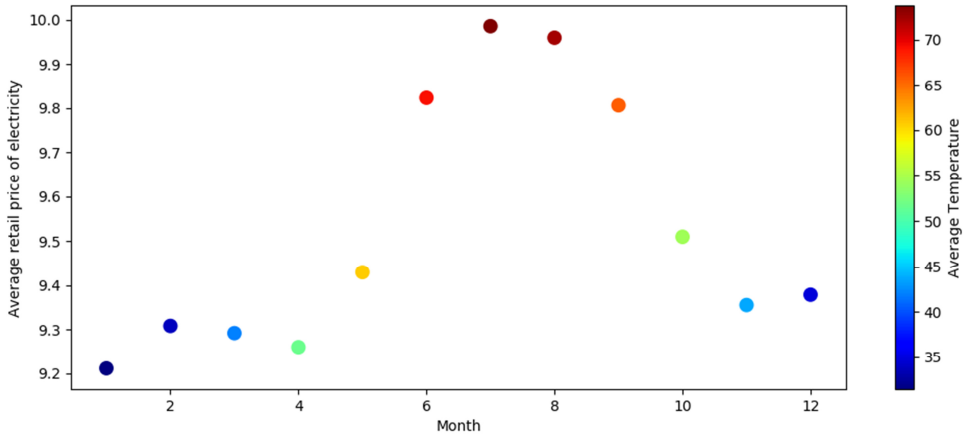
Figure 2. USA Annual Aggregated Average of retail price of Electricity



Relationship between Electricity price and Temperature¶

Aggregated data from 2001 to 2018 of average electricity price in USA per month is presented below. The months of higher temperature present higher prices. A seasonality or correlation between Electricity Price and Temperature is identified

Figure 3. USA Monthly aggregated Average of retail price of Electricity (cents per kilowatt-hour)



Relationship between Electricity Demand (Retail sales of electricity), Electricity Net Generation, price and revenue¶

As Electricity is non-storageble, the difference between the Net Generation and Retail sales of electricity is an overhead which is lost. Therefore, the importance of studying these features (Figure 4) and comparing them with price and revenue (Figure 5).

Figure 4. USA Monthly Aggregated balance between generated and sales electricity

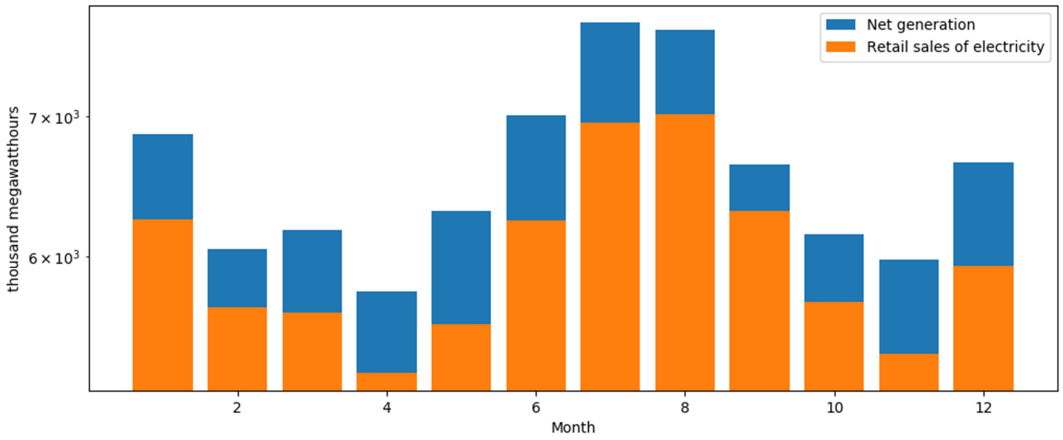
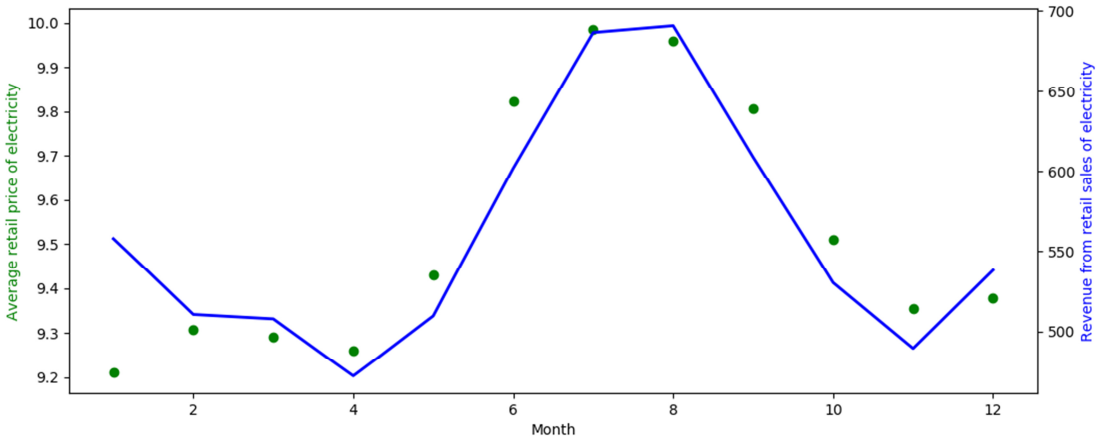


Figure 5. USA Monthly Aggregated average of Electricity Price (cents per kilowatt-hour) and Revenue ($Millions)



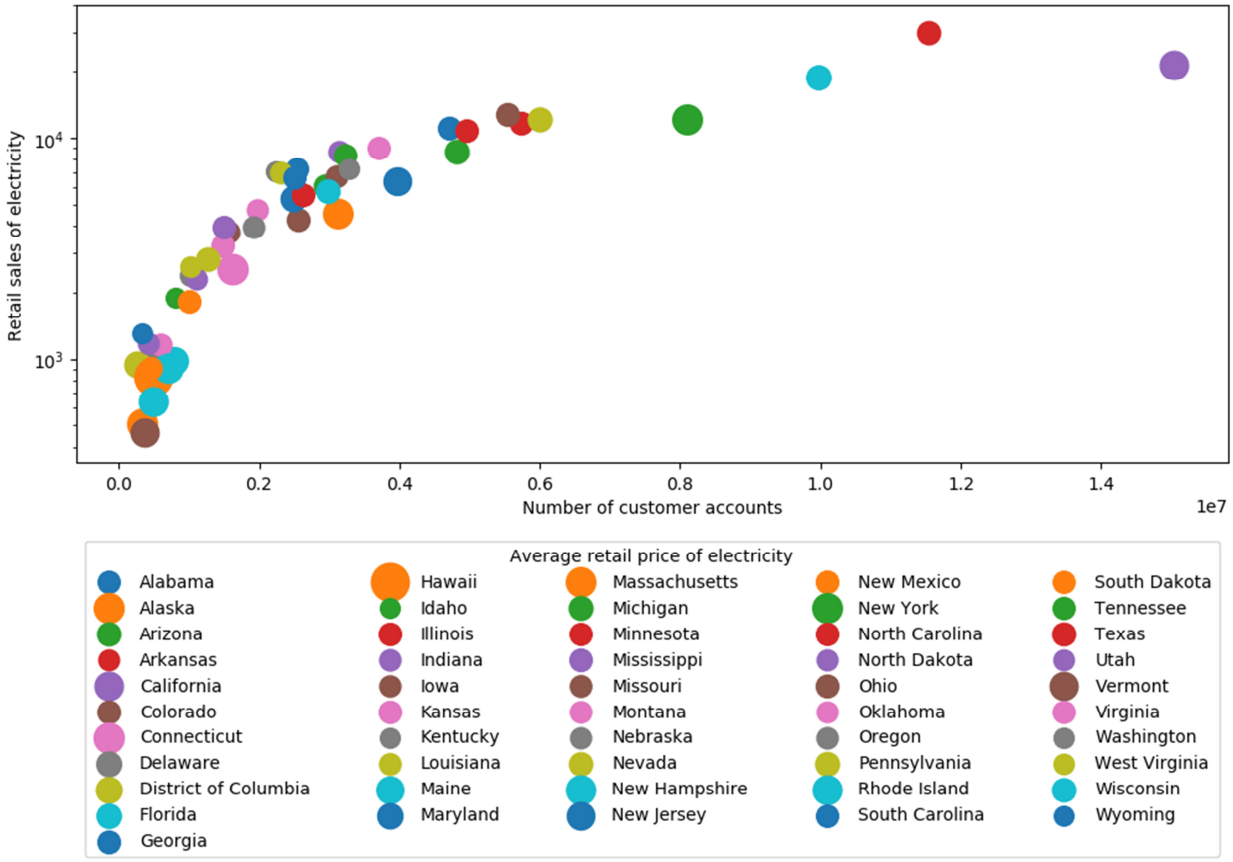
Looking at the two graphs above, we can appreciate:
1.- Revenue is higher when prices are higher.
2.- Consecutive months with similar temperatures/demands as Dec-Jan and July-Aug, the prices go down a bit the second month but revenue stay or increase.
Conclusion: We can infer that appropriate demand prediction lead to lower prices and higher revenue.

Relationship between Number of customer Account and Electricity Demand (Retail sales of electricity)¶

As expected, Figure 6 shows a relationship between the number of customer accounts and demand (Retail sales of Electricity).

The size of the bubble represents the aggregated average of retail price of electricity. At this point it is not clear the relationship between number of customer accounts and price, if any. But certainly there is a relationship between number of accounts and Net generation. Therefore, number of accounts is a feature that has to be considered in forecasting.

Figure 6. Annual Aggregated Average of Retail sales of Electricity (million kilowatt-hours) and Number of Customer Accounts

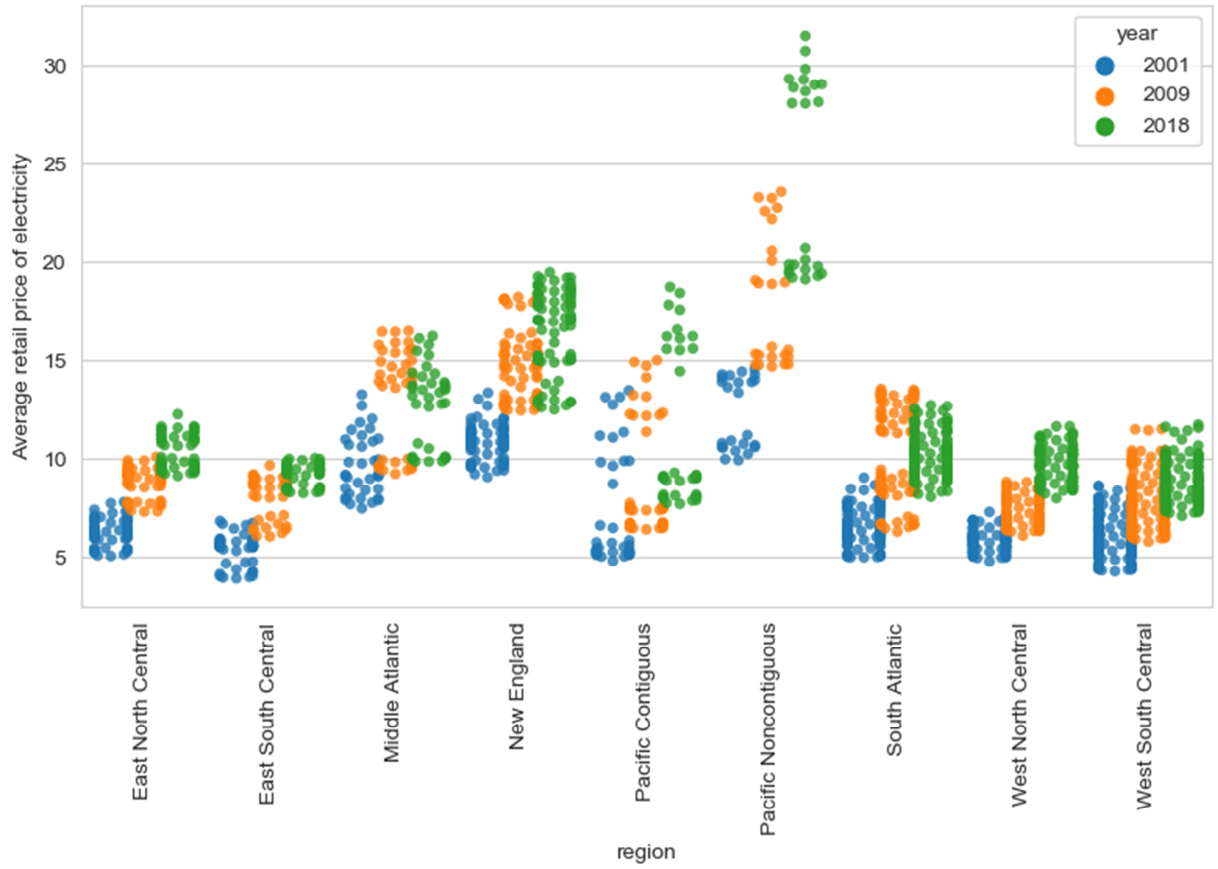


5. Exploratory Data Analysis (Inferential Statistics)

In Figure 1Figure 2 was visualized that the average retail price of electricity has increased over the last years. A deeper analysis is done by comparing data from three different years into the period 2001-2018 per region. For this analysis the years 2001, 2009 and 2018 were selected for comparison purposes. The figure below shows the percentiles (boxplot) and distribution of the average retail price of electricity per region.

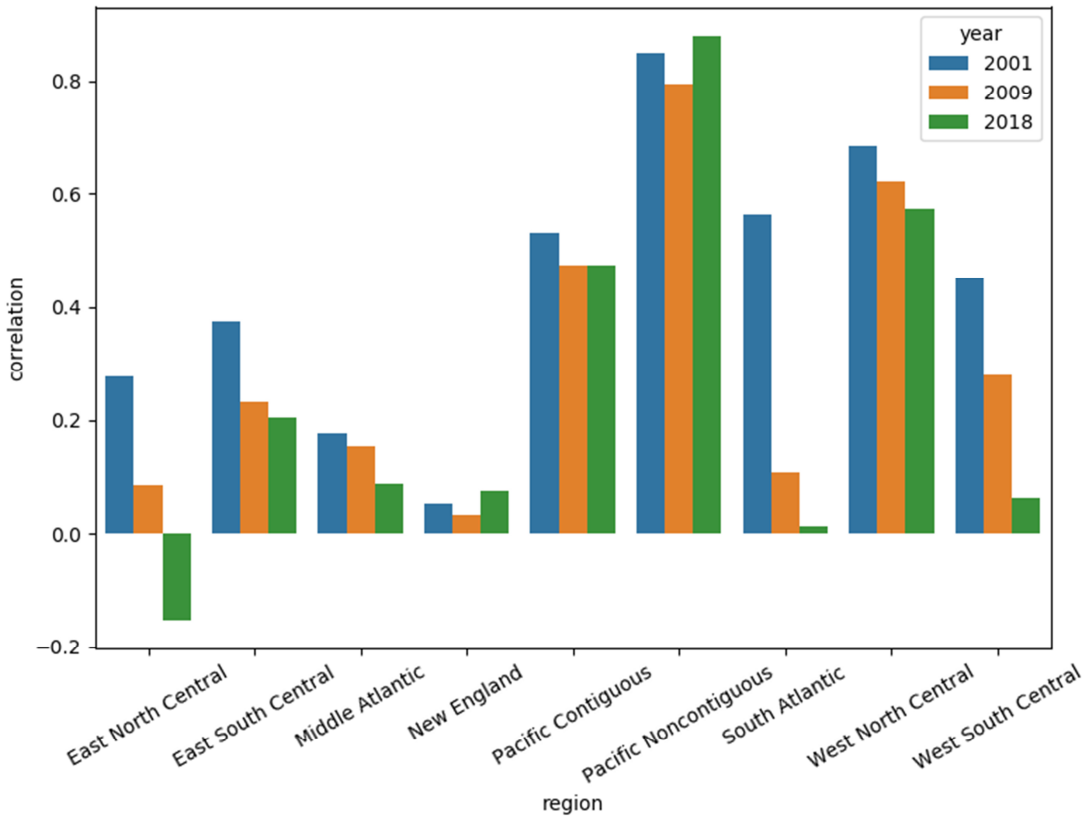
It is observed that there is a tendency in electricity price increase. No significant outliers were identified. The regions with higher average price variance are the pacific continuous and non-continuous regions.

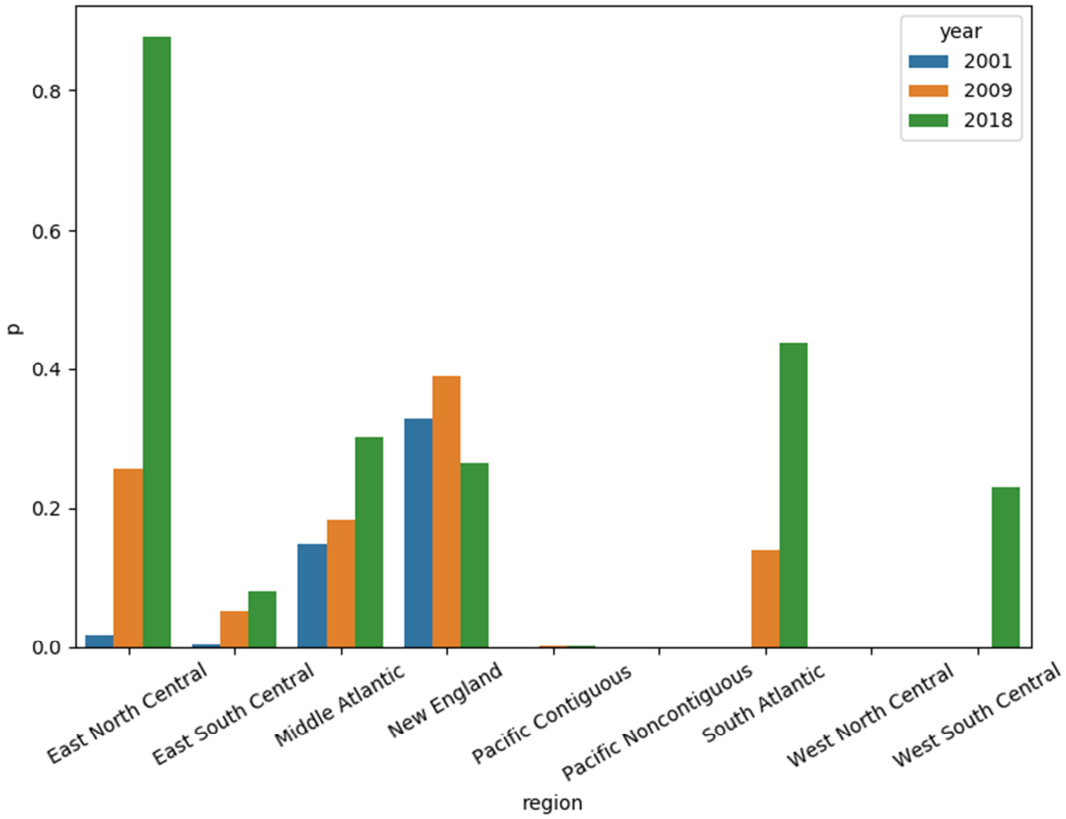
Figure 7. Analysis of Average Retail Price of Electricity by Region



A correlation between average temperature and price was visualized in Figure 3. Therefore a fast calculation of the aggregated USA temperatures (TMAX, TAVG and TMIN) was done, finding that TMIN (Minimum temperature) has the higher correlation. In this part, the correlation between TMIN and average retail price of electricity was quantified, as well as, the probability that such correlation might appear as shown below. The probability was found by bootstrapping the data per region per year.

Figure 8. Correlation (TMIN and Average Retail Price) and probability of occurrence of such correlation per region





The findings indicate that low correlation between TMIN and Retail Price is something that is likely to happen but high correlation does not seem to be very likely.

6. Machine Learning

The Data is Time Series and non-stationary due to:
1.Trend - the Retail price of Electricity grows over time. The trend component of our predicted variable might be due to inflation or some other macro-economic factors that are not reflected in the collected features into the data.
2.- Seasonality - there is a periodic change of the price which might be considered in some models as sesonality. Nevertheless, It was studied that the Retail price of Electricy tend to be higher in months with higher temperature. Thus, depending on the model these changes can be considered as a consequece of the correlation with temperature and/or other features.

Models under consideration¶

According to the observed in the Project1_Part3 - Statistic Inference, it would be necessary a model per region. Nevertheless, for the initial model selection,the predicted variable would be the US average temperature in order to simplify complexity and reduce computation time.
Two kind of models were considered: Time-series univariable models and multivariable models.
Time-series univariable models:

1. Moving Average
2. Facebook profet
3. ARIMA

These kinds of time series uni-variable models base their predictions only on the date-time information. The above considered models are traditionally used for time-series predictions.
Multivariable models:

1. Moving Average + Linear Regression (Combined model)
2. Moving Average + Lasso (Combined model)
3. Moving Average + Ridge (Combined model)
4. Moving Average + ElasticNet (Combined model)
5. Long Short-Term Memory (LSTM) (Unsupervised model)

Combined models - Time-series univariable model + atemporal multivariable model: it considers a Time-series univariable model to forecast the trend + an atemporal multivariable model to forecast the differentiated data. The intention is to remove the trend part (inflation and other macro-ecomic factors) by using moving average and predict the remaining part (which depends on temperature, net generation, consumption, etc) by using different regressors. Thus, the predicted value would be the addition of both predictions.

Metrics for model Selection¶

The metrics for model selection are:

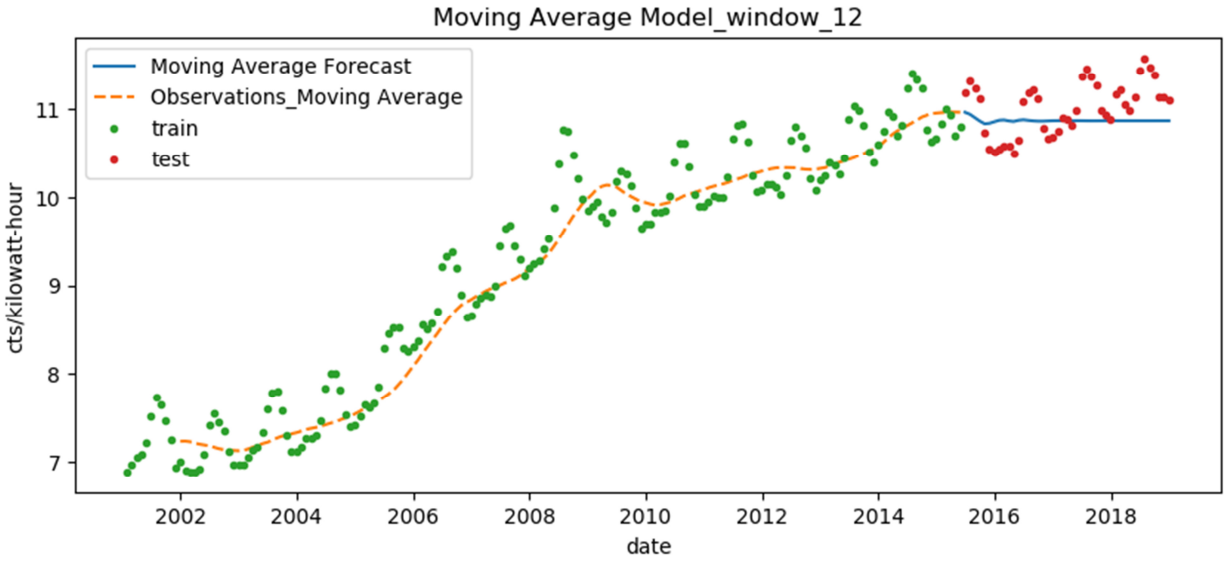
• rmse (Root-mean-squared-error), and rmsloge (Root-mean-squared-logarithm-error) for a test size of 0.2 • cv_mean_rmse: mean of rmse of the cross validations of the tuned models by using TimeSeriesDataSplit of ScikitLearn with a number of time splits of 7

Findings¶

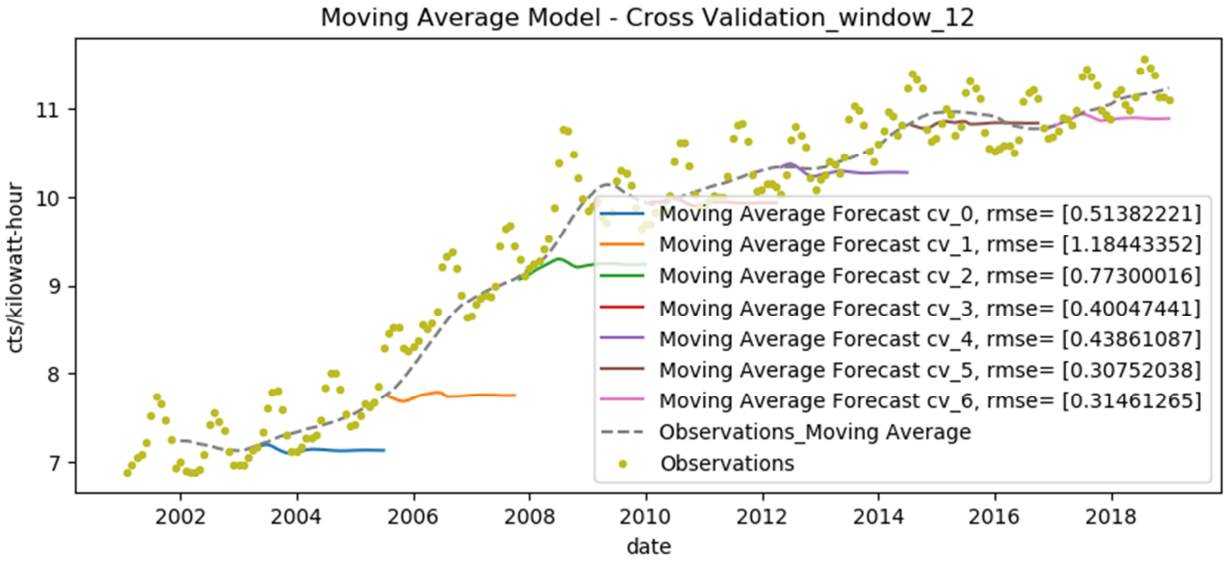
Moving Average¶

(Time-series univariable model)

As shown below, moving average smooth out short-term fluctuations and highlight longer-term trends or cycles. Therefore, moving average forecasting is not the best choice when looking for price fluctuations in short term but a decent option for trend forecast.



rmse: 0.32305920425052803
rmsloge: 0.026840902065889544



cv_rmse_mean: 0.5617820284708822
cv_rmsloge_mean: 0.05471147103489592

Facebook Prophet (fbprophet) also known as Prophet¶

Time-series univariable model

Reference: https://facebook.github.io/prophet/docs/seasonality,_holiday_effects,_and_regressors.html

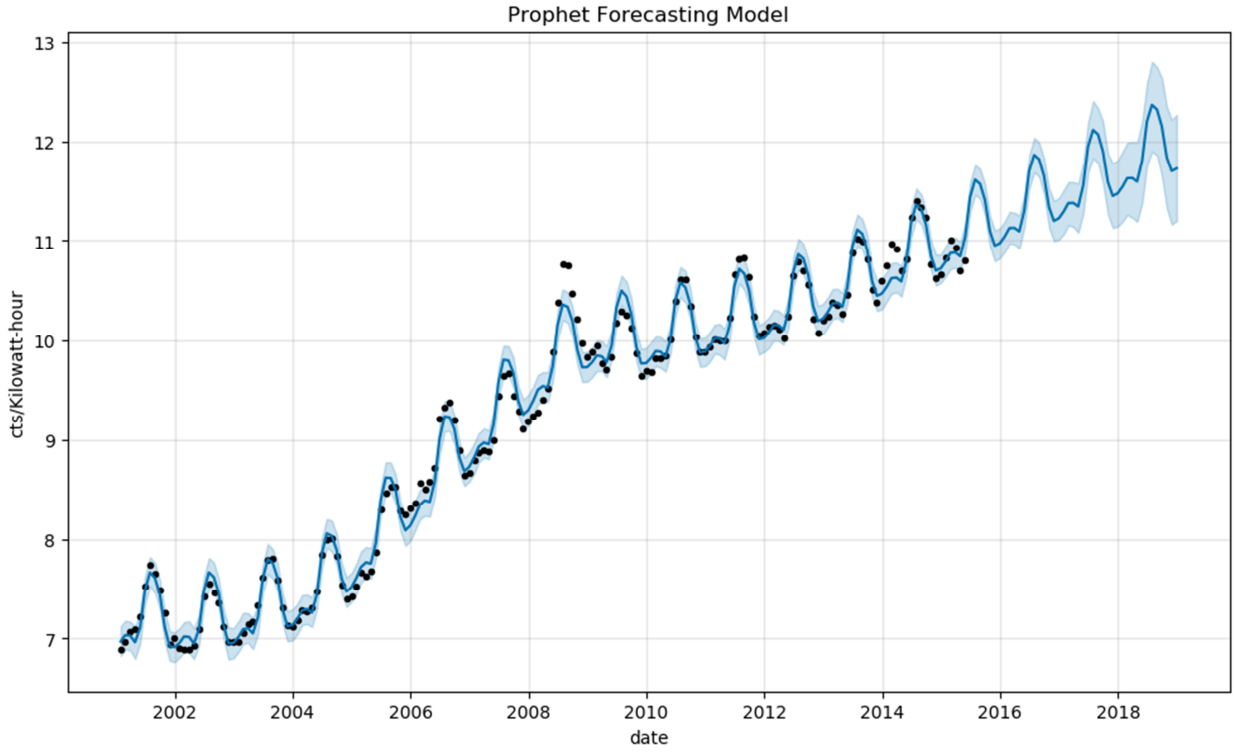
Prophet is a univariate model. It could be considered a naive model since it intends to predict the outcome based just on the dates (time series). The model descompount the outcome/response variable *Y* behavior into trend and seasonality parts.

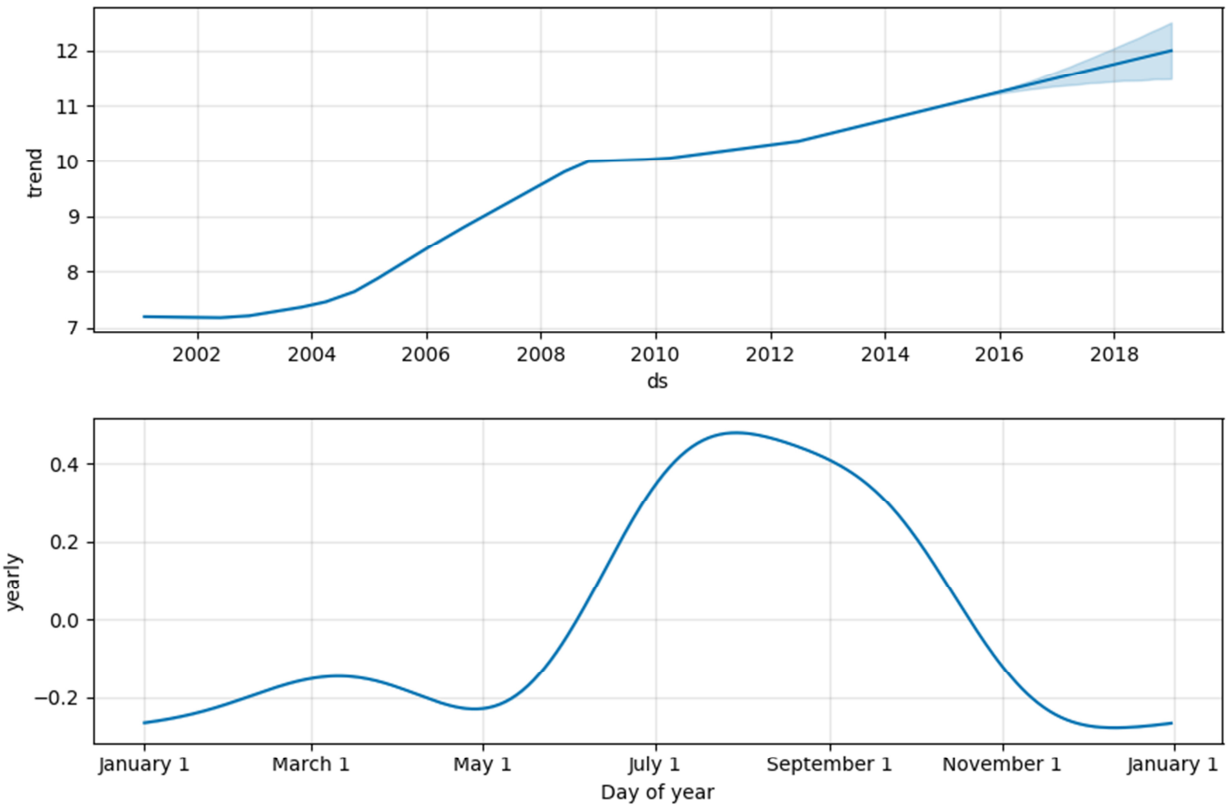
Y =Xi(1) Trend + Xi(2) Seasonality

The trend estimator uses linear regression with changing points as hyperparameter. It was used 'autoscale' which is the default hyperparameter was used.

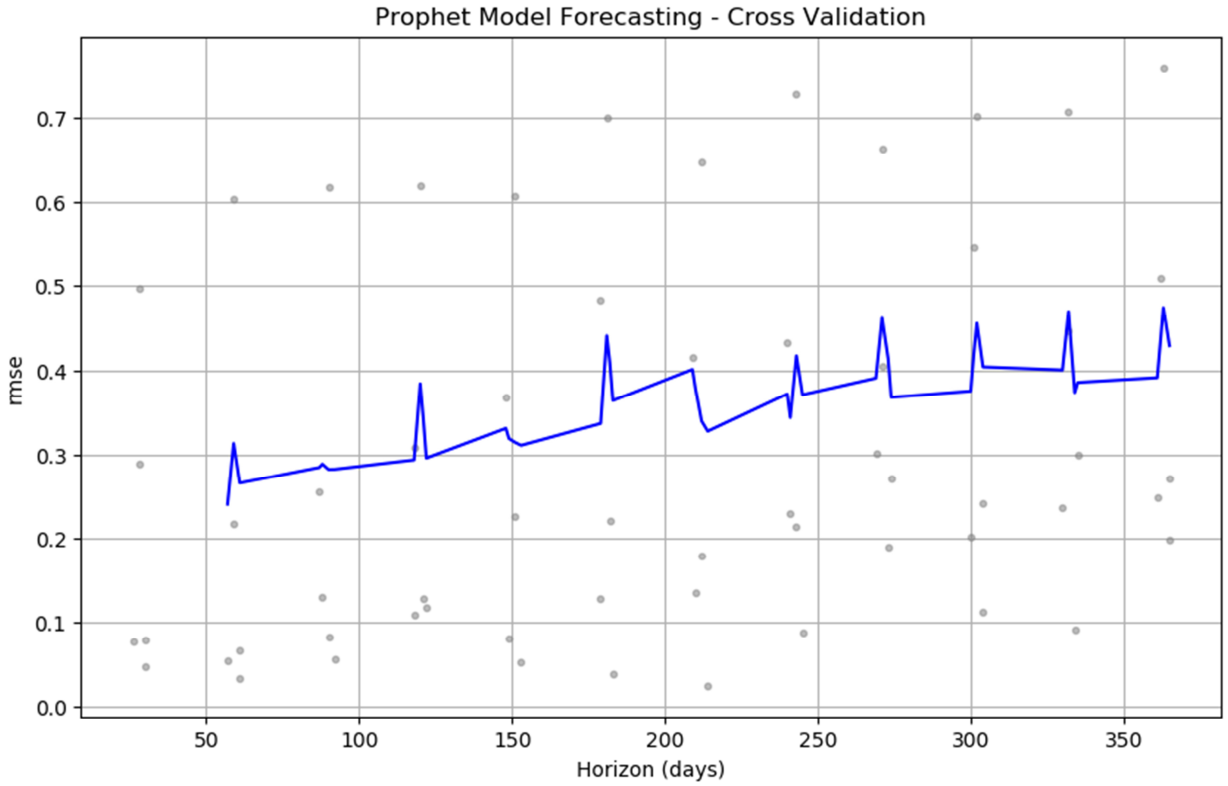
Seasonalities are estimated using a partial Fourier sum, which is a well known mathematic method traditionally used to model periodic electrical signals. The parameter fourier_order=5 was used. This values was selected by inspection of the graphs of the components since in Project 1_part2 was visualized the dependency of variation of price with respect month of the year due to temperatures. Also, an additive annual seasonality of 365 days was added based on the data analysis done in previous parts.

According to the cross-validation of fbprophet model graph below, the error increases when the test size or prediction window (Horizon) is bigger which an expected result is.





test_size: 0.2
rmse: 0.5737011447405002
rmsloge: 0.04641876087501963



cv_rmse_mean:0.36471317748245685

ARIMA (autoregressive integrated moving average)¶

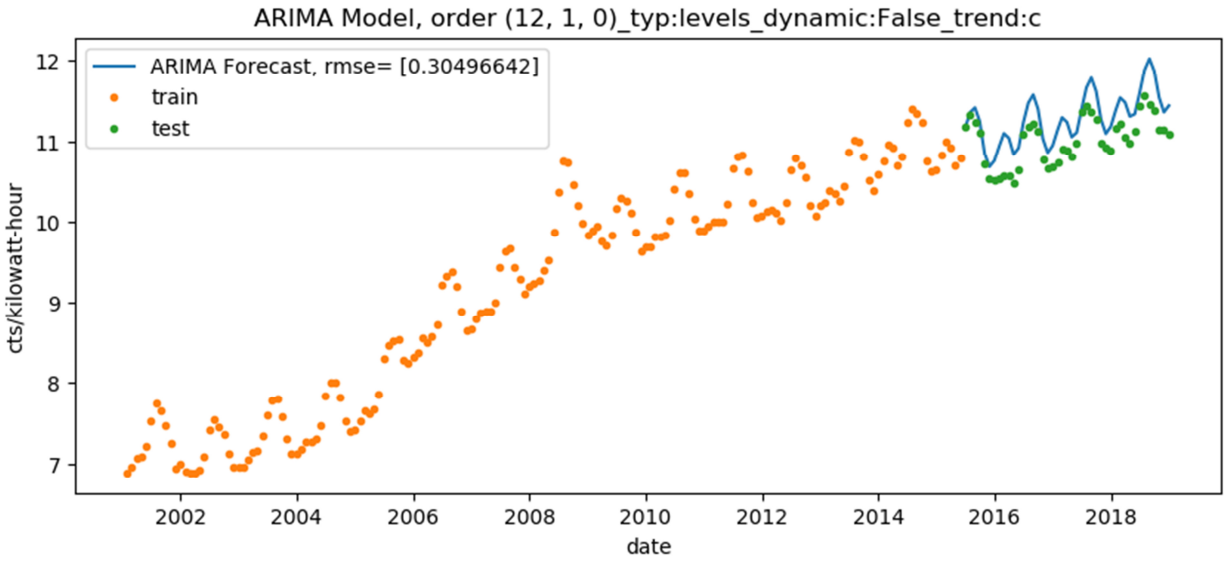
Time-series univariable model

Refrence: http://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.fit.html

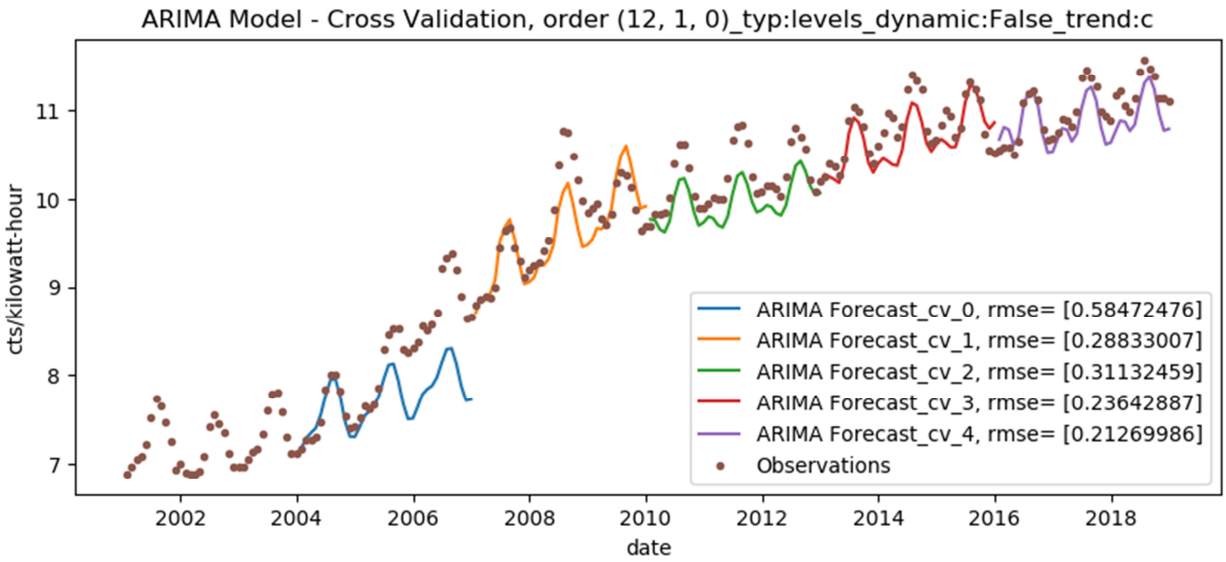
GridSearch from scikit-learn could not be used with ARIMA estimator as it does not implement a 'get_params' methods. Thus, a custom Gridsearch with cross validation was performed. It was found that the best ARIMA model hyperparameters are: order (12,1,0), typ=Levels, Trend=constant, Dinamic=False.

Due to the date dependency, the data split (done by TimeSeriesDataSplit of ScikitLearn) is performed sequentially taking different chunk sizes of sequential data. Bigger the number of splits, smaller the initial training sets. Therefore, there are additional convergence problems when the training data set is too small.

A manual cross-validation, in order to find the best tuned model, was done by splitting the data with TimeSeriesDataSplit of ScikitLearn. At most of the cases, the performance metrics (rmse and rmsloge) seems to be consistent with exception of the cv_0 which might be due to the short size of the training set.



rmse: 0.3049664198267251
rmsloge: 0.025039997481575704

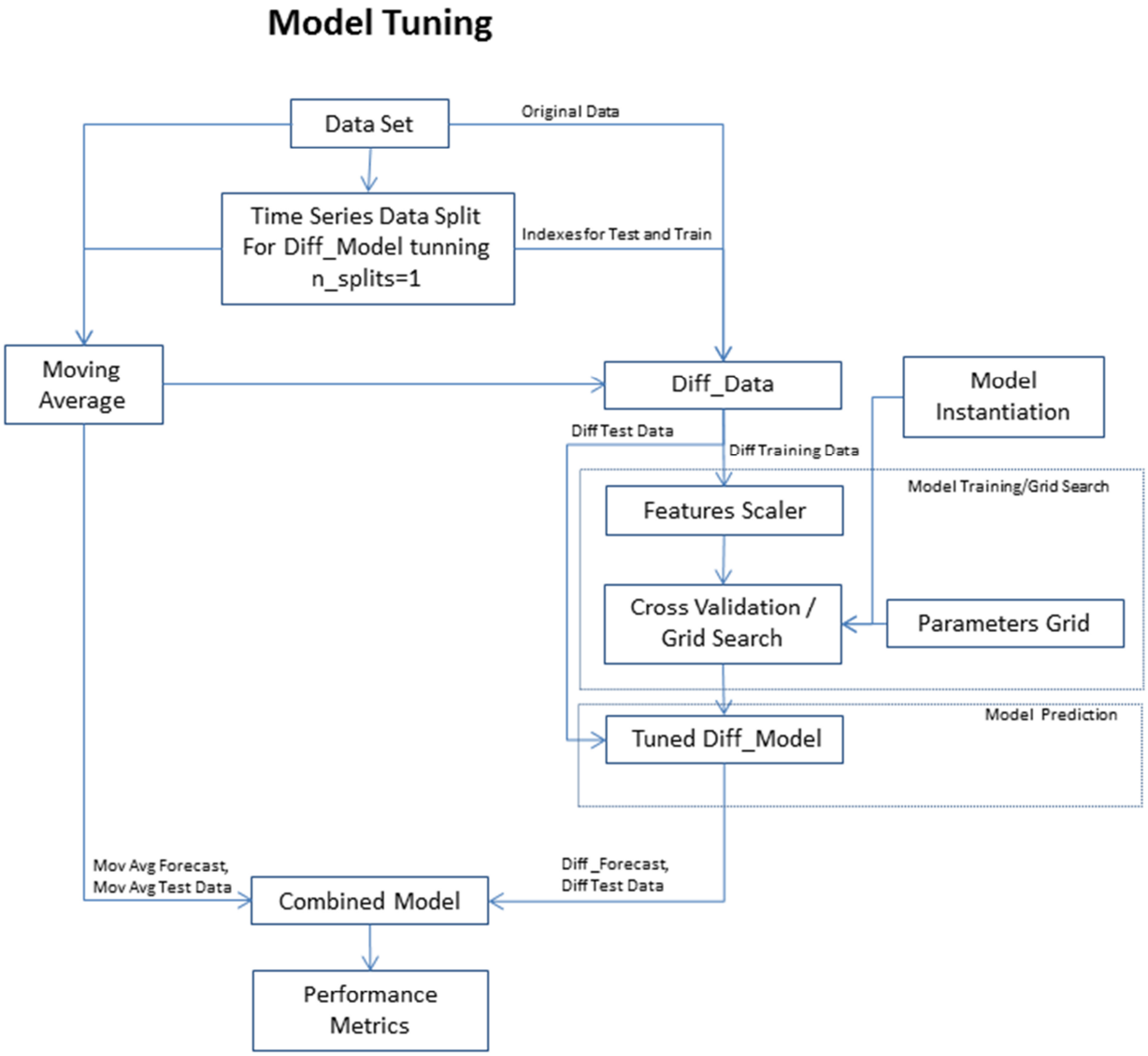


cv_rmse_mean: 0.32670162859826984
cv_rmsloge_mean: 0.03083668305626222

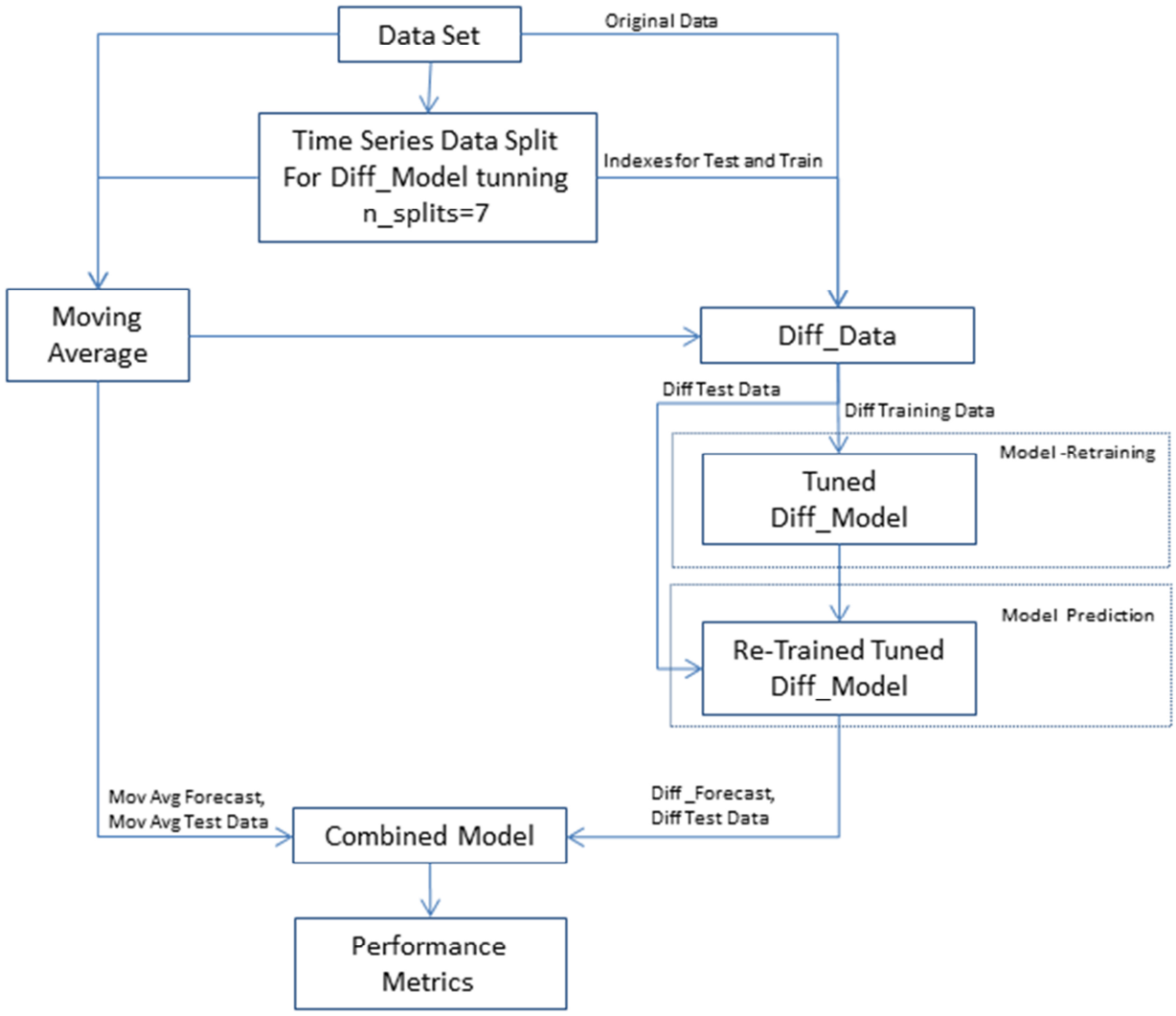
Moving Average + Other regressors¶

In this part, the moving average is calculated to estimate the trend and substract it from the outcome/response variable *Y* in order to eliminate the growing trend. The differentiated outcome/response variable diff_*Y* without trend is then considered an atemporal data set. The atemporability is giving by the fact that the "seasonality" changes can be predicted by using the features that affect such changes others than date/time.
Also, due to the scale difference between different features, pipeline consistent of a Scaler and a model from ScikitLearn had to be applied.

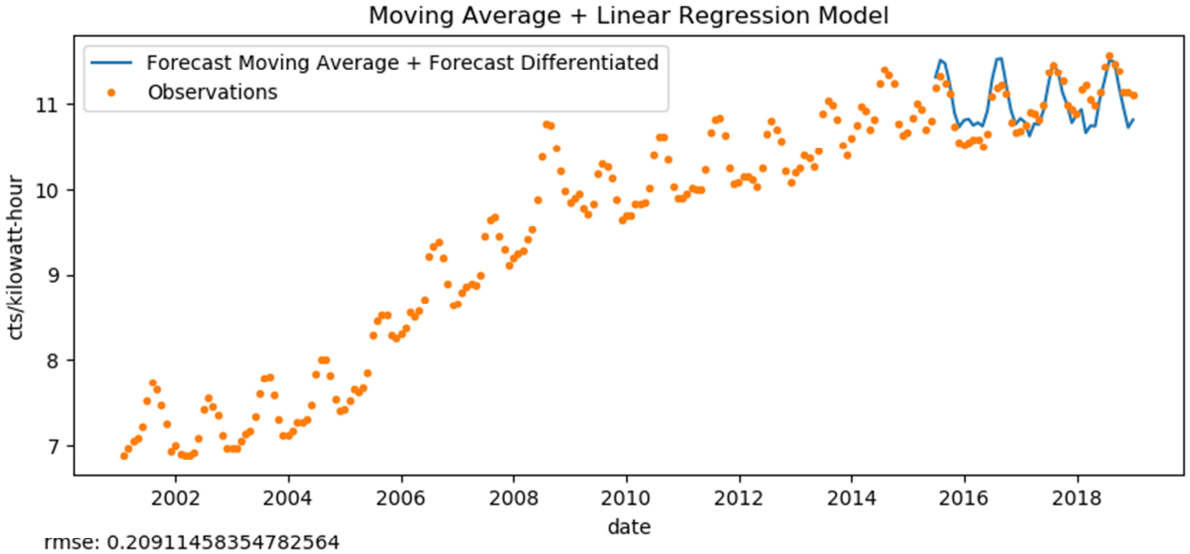
The standard cross-validation of the GridSearch from ScikitLearn was applied in this section of model tuning since the differentiated data is considered at this point atemporal.
Once, the best fit-parameters were found for the model of the differentiated data (See model Tuning diagram), It was found that passing just the best parameters to the 'differentiated' model to train it with smaller training data set was not as good as when the model is re-trainned with a small data set. Therefore, the model is saved to be re-trained for later predictions (See Cross-validation model assessment).



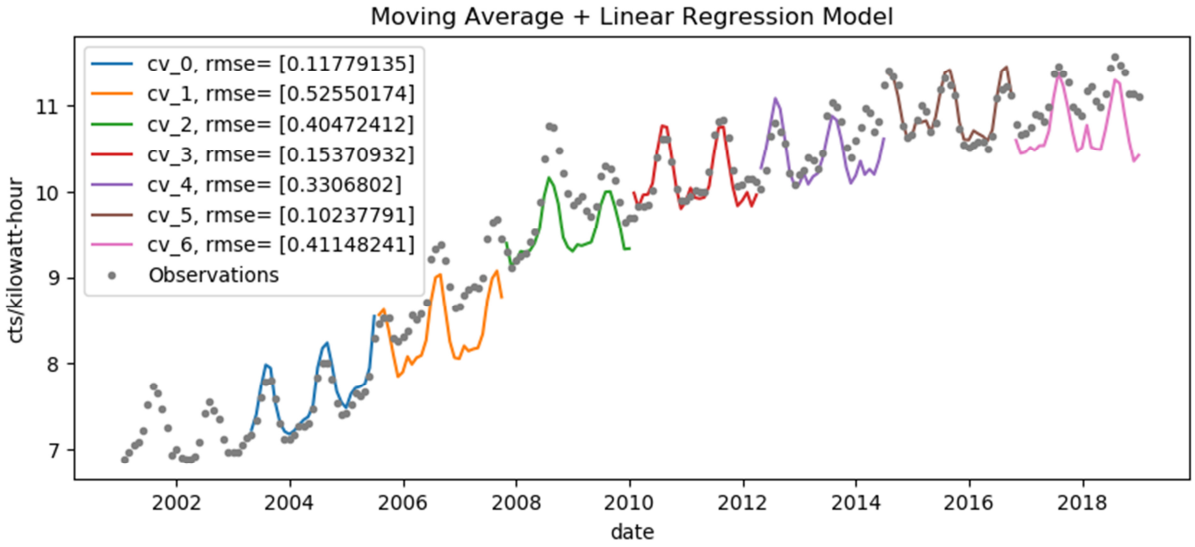
Cross Validation – Model Assessment



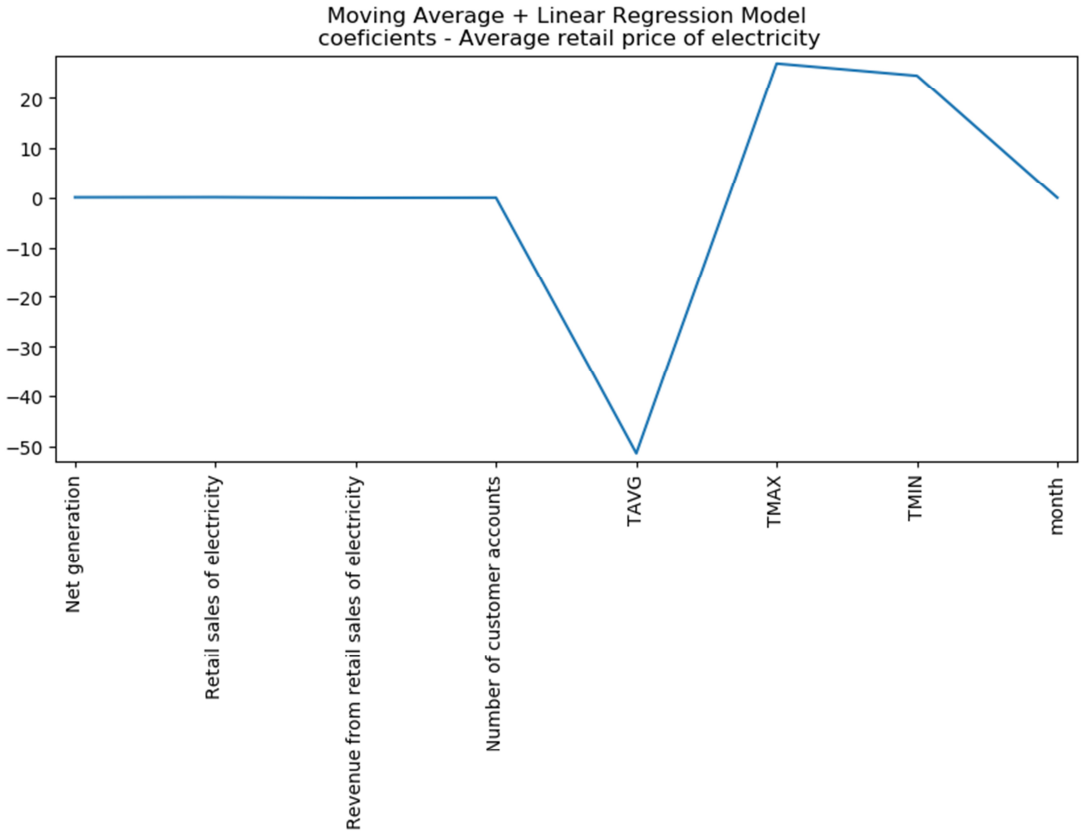
- Moving Average + Linear Regresor **



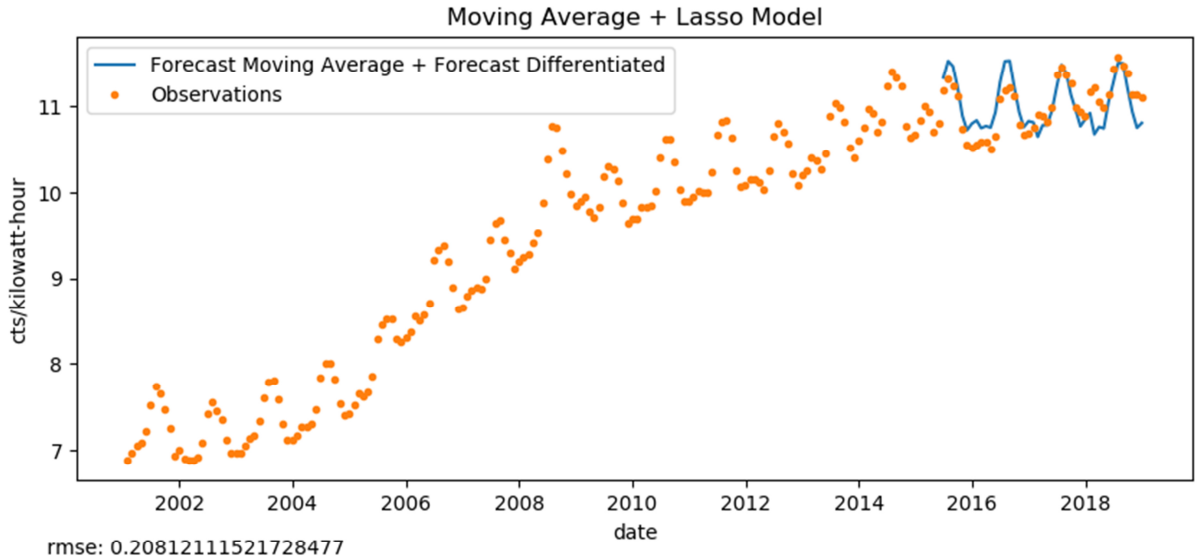
rmse: 0.20911458354782564
rmsloge: 0.0174835179459242
best_fit_parameters: {'model__fit_intercept': True}



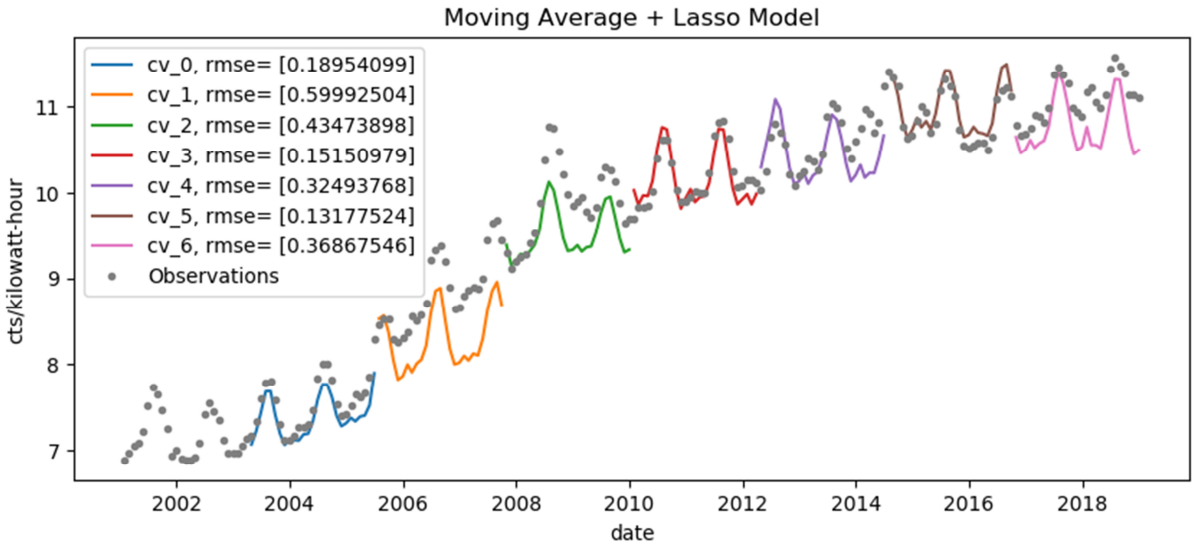
cv_rmse_mean: 0.2923238648976203
cv_rmsloge_mean: 0.027208356374431922



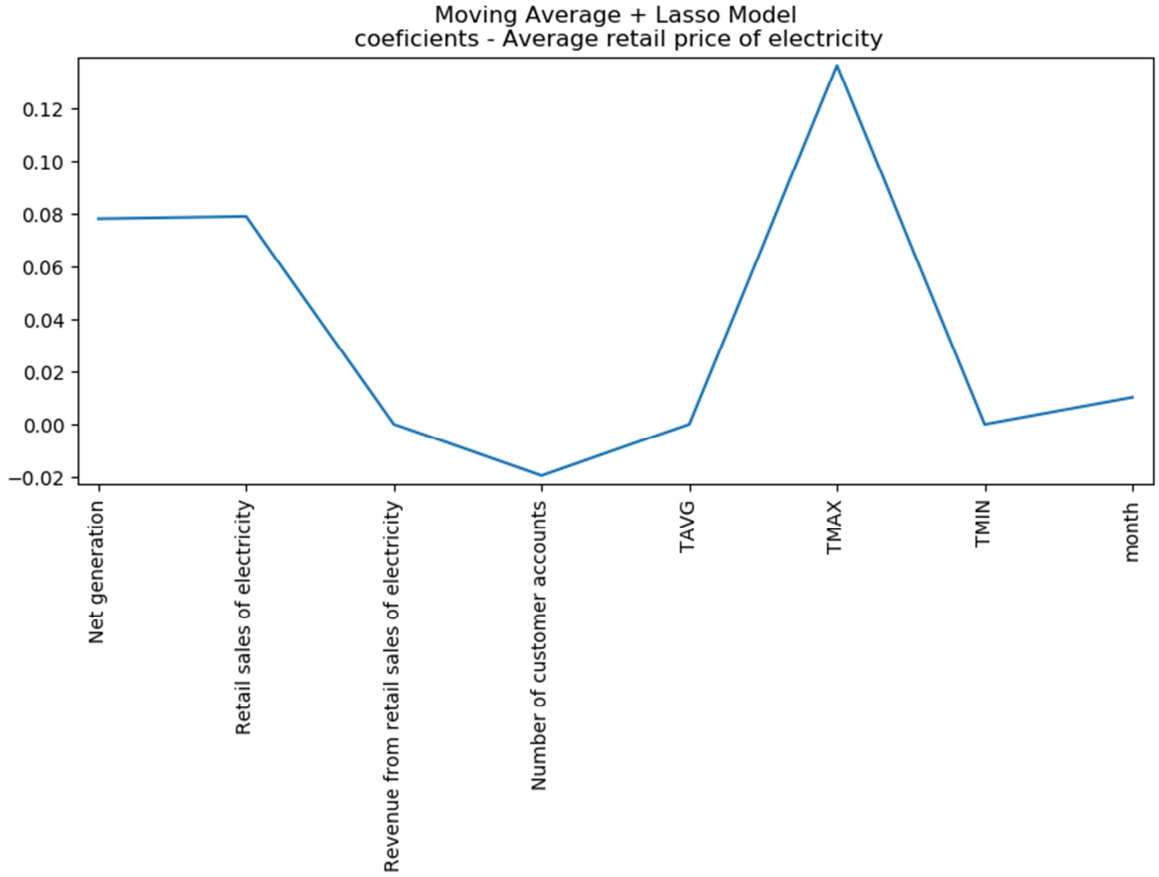
- Moving Average + Lasso Regresor **



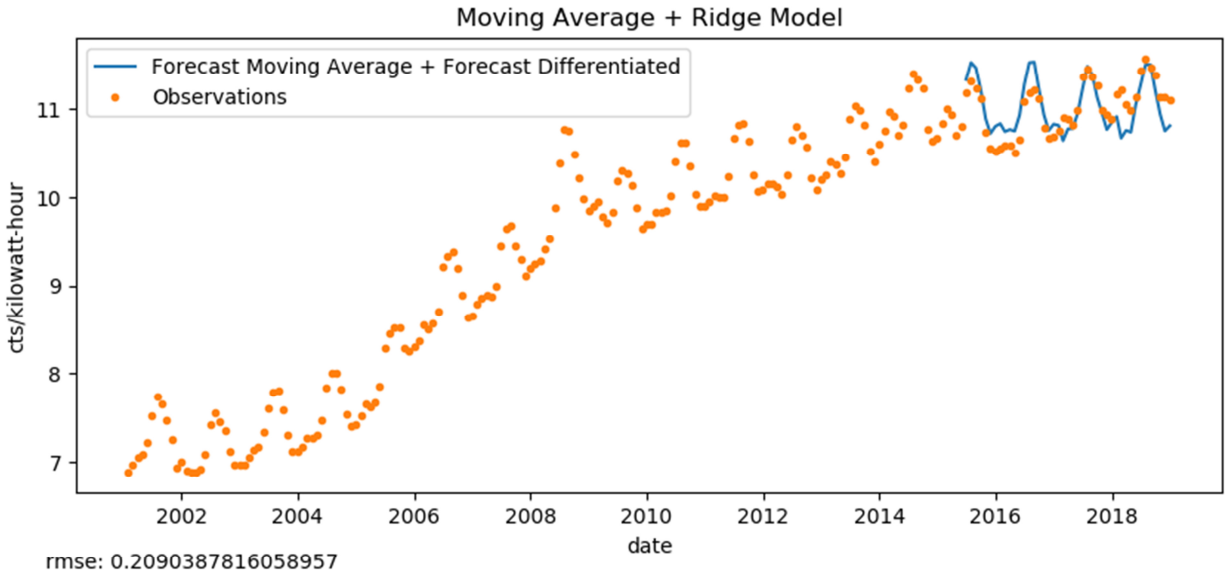
rmse: 0.20812111521728477
rmsloge: 0.017392478235886705
best_fit_parameters: {'model__alpha': 0.0020235896477251557}



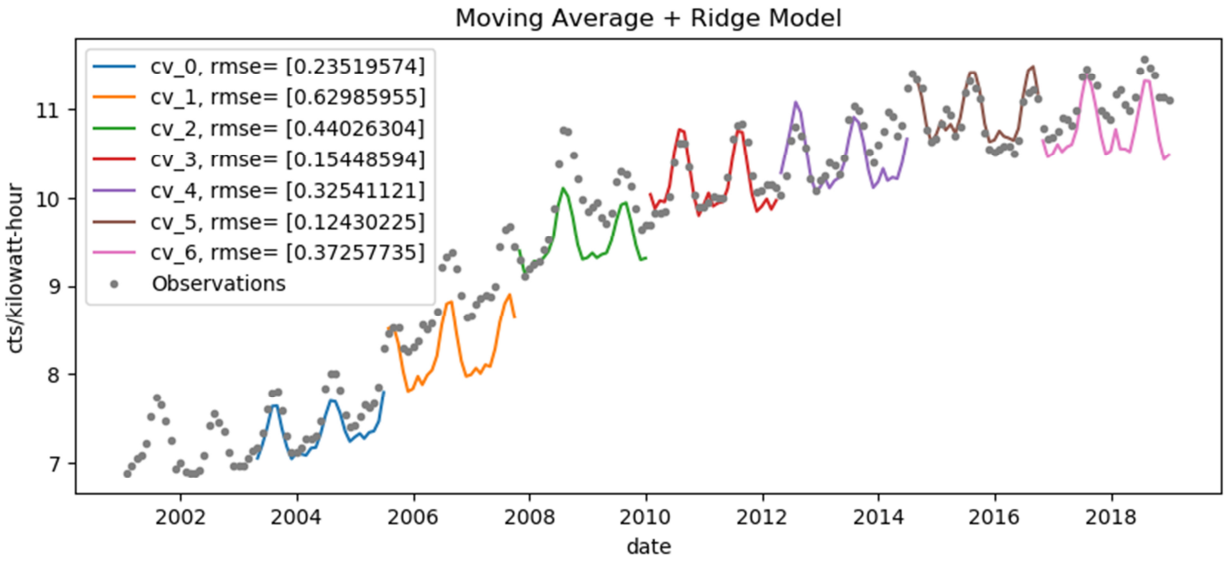
cv_rmse_mean: 0.31444331092868694
cv_rmsloge_mean: 0.029685506172999986



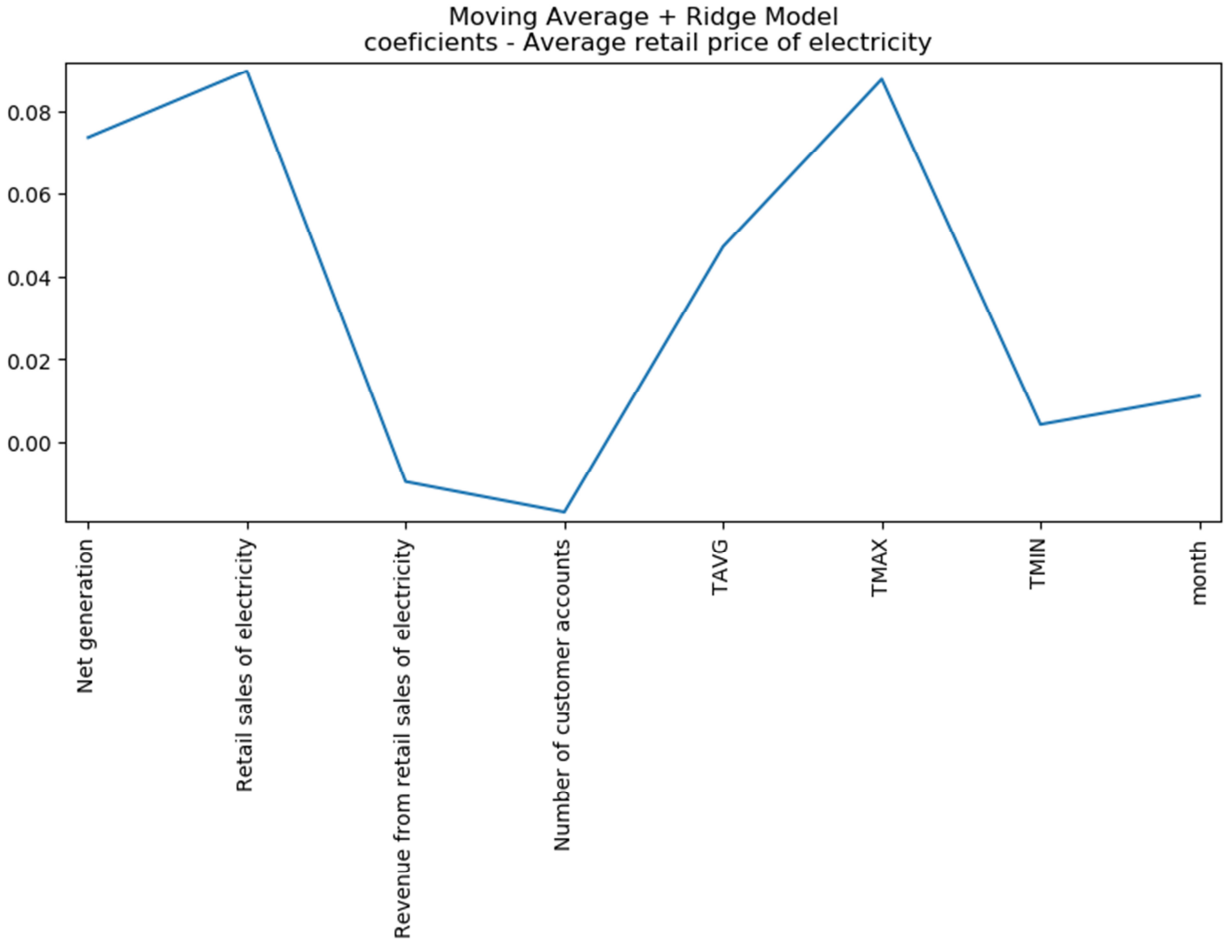
- Moving Average + Ridge **



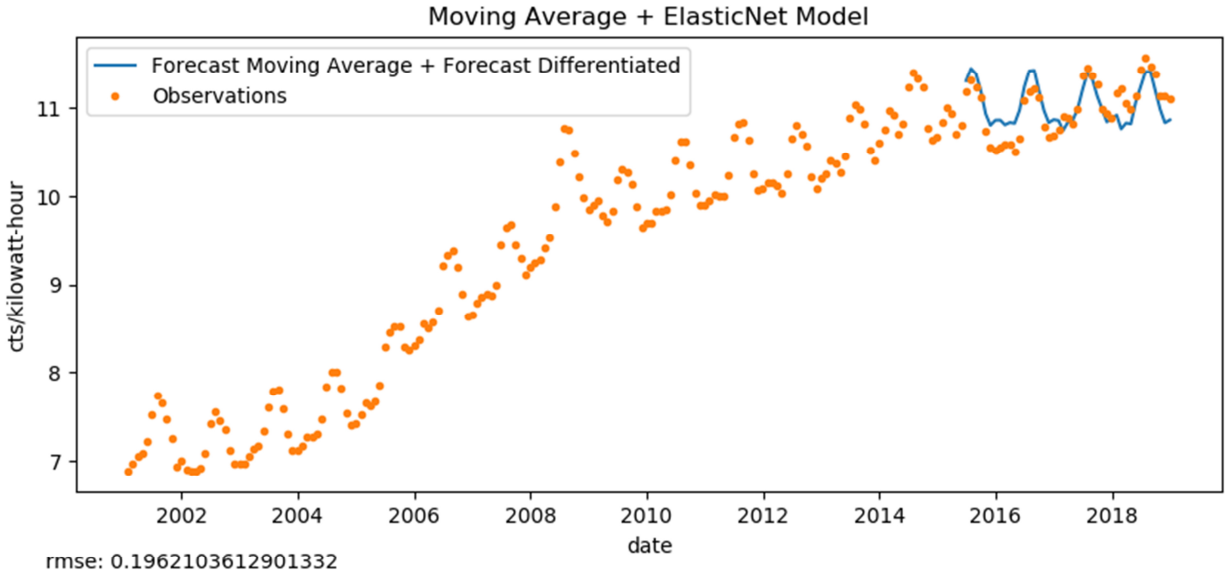
rmse: 0.2090387816058957
rmsloge: 0.017468730006212987
best_fit_parameters: {'model__alpha': 1.0, 'model__fit_intercept': True}



cv_rmse_mean: 0.3260135838894967
cv_rmsloge_mean: 0.030976316579231916

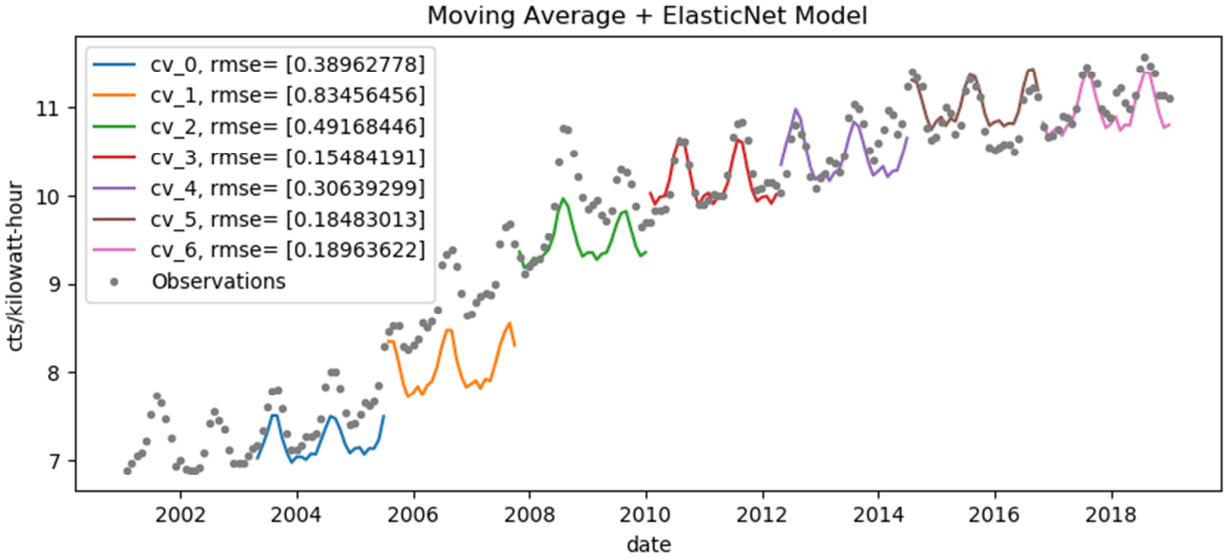


- Moving Average + ElasticiNet **

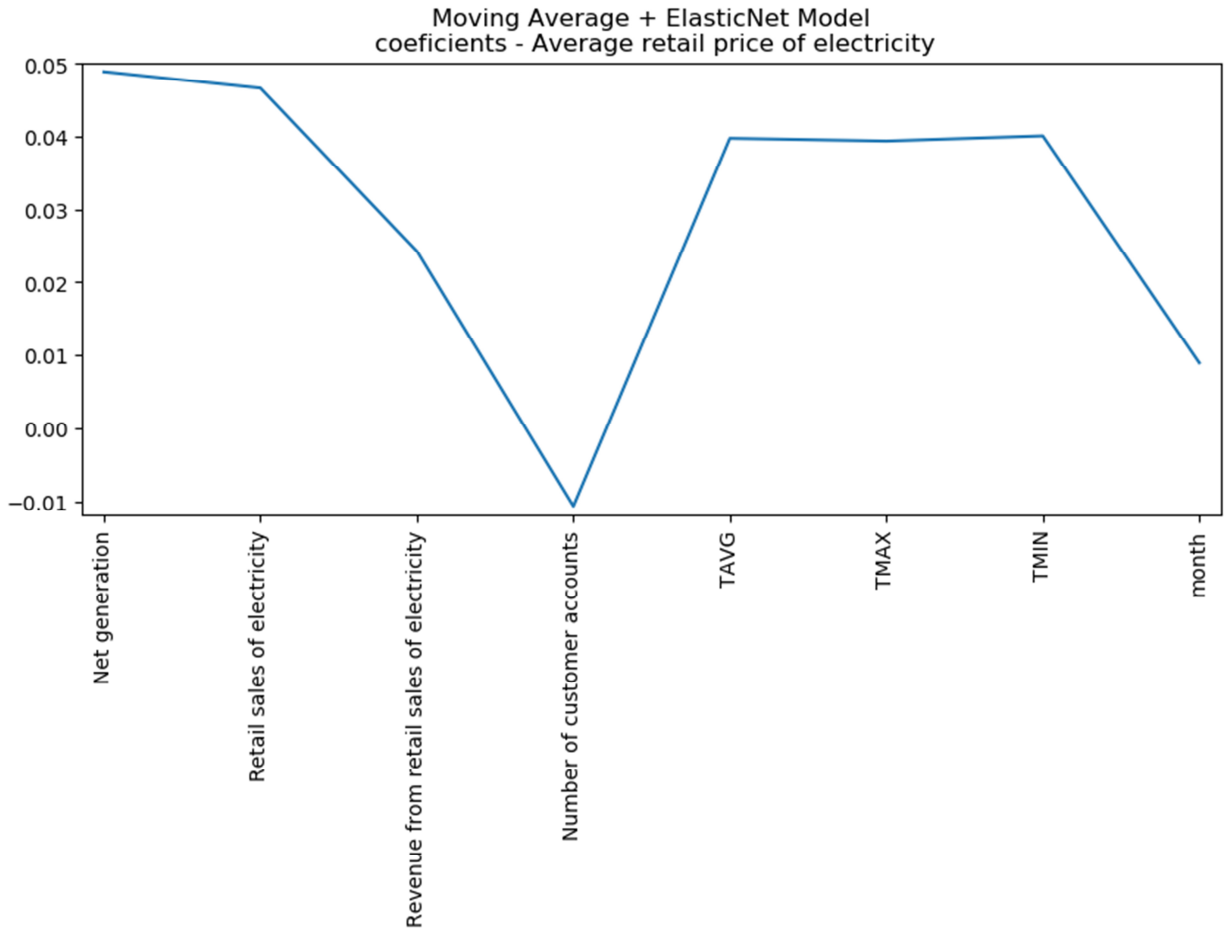


rmse: 0.1962103612901332
rmsloge: 0.016456531690803453
best_fit_parameters: {'model__l1_ratio': 0.0}

-



cv_rmse_mean: 0.3645111519213976
cv_rmsloge_mean: 0.035670365450674556



7. Model Selection Metrics Summary

	cv_rmse_mean	cv_rmsloge_mean	rmse	rmsloge
Moving Average	0.5618	0.0547	0.3231	0.0268
fbprophet	0.3647	-	0.5737	0.0464
Mov_Avg+Linear Regression	0.2923	0.0272	0.2091	0.0175
Mov_Avg+Lasso	0.3144	0.0297	0.2081	0.0174
Mov_Avg+Ridge	0.3260	0.0310	0.2090	0.0175
Mov_Avg+ElasticNet	0.3645	0.0357	0.1962	0.0165
ARIMA	0.3267	0.0308	0.3050	0.0250

8. Conclusions

- Time Series models as fbprophet and ARIMA are good options to forecast prices based on historical trends and seasonality. As fbprophet breaks down the seasonality and trend, this model also allows forecasting the fluctuations that actually impact the electricity business operations.
- It is possible to break down the trend and the seasonality using the combined models Mov_Avg + other regressor. The other regressors as Linear Regression, Lasso, Ridge and ElasticiNet have performance comparable to fbprophet and ARIMA model. The main difference is that the time series models as fbprophet and ARIMA use as only feature the date-time information while the combine models use other features to forecast the differentiated (seasonality) part.
- The combined model Mov_Avg + Linear Regression uses as features the temperature information (TAVG, TMAX, TMIN) while the other considered combined models consider Net Generation, Retails sales of Electricity, Revenue, Number of customer accounts and month beside the temperature features.
- The best model to use would depend on the purpose of the forecast and the available information.
- It is recommended to re-trained the selected model with new data in order to capture changes on patterns of behavior.

9. Next Steps

The current analysis was based on US average temperature. Nevertheless, individual models per US region has to be tuned in order to be able to forecast Electricity price at any state of USA.

10. Code & other materials

Code:

Project1_Part1. Data Acquisition and Data Wrangling

https://github.com/megiza/Project_ElectricityRates/blob/master/Project1_Part1.ipynb

Project1_Part2. Data Visualization

https://github.com/megiza/Project_ElectricityRates/blob/master/Project%201_Part2.ipynb

Project1_Part3. Exploratory Data Analysis

https://github.com/megiza/Project_ElectricityRates/blob/master/Project1_Part3.ipynb

Project1_Part4. Exploratory Data Analysis

http://localhost:8888/notebooks/Documents/Python/Springboard/CourseCurriculum/FirstCapstoneProject/Project1_Part4.ipynb

Presentation:

Presentation_Electricity_rates

https://github.com/megiza/Project_ElectricityRates/blob/master/Presentation_Electricity_Price_prediction.pdf