Project Technical Report

Group 6: Shumin Song, Frank Li, Qifan Jiang

I. Introduction

This project is aiming at predicting the amount of electricity consumption for various clients in different lengths of time interval by utilizing historical data “Electricity Load Diagrams 2011-2014 Data Set” from UCI Machine Learning Repository and time series machine learning techniques such as the SARIMAX model.

The primary purpose of our group is analyzing monthly electricity consumption in specific(precise amounts for different types of clients) perspectives. For example, some client groups might have higher consumption than others in a specific time period. In order to get a more accurate consumption forecast, we will train monthly consumption data for a representative group of clients and plot a predicting graph to show monthly change of test group’s consumption. We will plot the precise consumption for particular types of clients on the same graph to gain direct, clear comparison. Based on this analysis, we want to predict the monthly amount of electricity consumption in the following year 2015 and its effectiveness will be measured using mean absolute percentage error.

These predictions can help power companies plan their energy production, distribution, and pricing strategies more effectively and consumers make informed decisions about their energy consumption habits.

II. Data extraction and processing

Our data source is “Electricity Load Diagrams 2011-2014 Data Set” from UCI Machine Learning Repository. Here is the URL of original data set:<https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014#> . The original data set is documented in a “.txt” file, which has semicolon “;” as delimiter. There are 31 variables in total. The first column represents date and time as a string with the following format “yyyy-mm-dd hh:mm:ss” while every record has a 15-minute gap between its previous and following record. Each remaining column represents the amount of electricity consumption of a client in kW during a 15-minute interval. The consumption amount is presented in float value. To avoid ambiguity in values, we substitute comma “,” in the original data set which represents decimal point into ordinary dot “.”.

The data set contains electricity consumption records from 2011-01-01 to 2015-01-01 and has 140,256 observations altogether. All time labels report to Portuguese hour. Since we want to predict consumption in 2015, the record of 2015-01-01 is discarded in our processing. There are no duplicate records, missing values or extreme outliers in this data set. Nevertheless, there are still some minor problems. Some clients were created after 2011 so their consumption in 2011 is considered zero. Therefore, our group will only use records from 2012 to 2014 to classify clients. In addition, every year in March time changes day (which has only 23 hours) the values between 1:00 am and 2:00 am are zero for all points. Every year in October time changes day (which has 25 hours) the values between 1:00 am and 2:00 am aggregate the consumption of two hours.

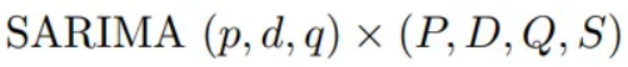
III. Variable Analysis

(1) Target Variables

In this project, we only focus on one target variable, which is the amount of electricity consumption in kW units. However, this target variable has different meaning in our two major project objectives. In the first objective, it represents daily and monthly average electricity consumption of all clients for general perspective. In the second objective, it represents monthly electricity consumption of a particular type of client.

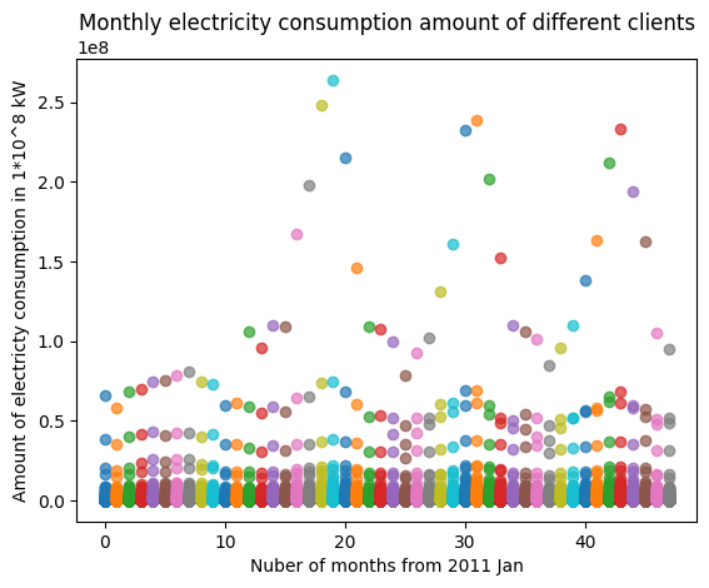
(2) Predictive Variables

The target variable is predicted by several variables in our algorithmic solutions. The SARIMAX(Seasonal Auto-Regressive Integrated Moving Average) model is the primary method we use to estimate the amount of electricity consumption with different time interval basis(hourly, daily, monthly). Here is a shorthand notation of SARIMAX model:



In this model, p = non-seasonal autoregressive (AR) order, d = non-seasonal differencing, q= non-seasonal moving average (MA) order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = length of repeating seasonal pattern. We will apply this model for not only the general average consumption analysis but also average consumption of each type of client. A particular type of client may also act as a coefficient in predicting the specific average electricity consumption.

(3) Clients classification



We accumulated each monthly electricity consumption amount from 2011 to 2014 and created a scatter plot to show the general trend and distribution of all clients. As this graph shows, there are five significant clusters for clients’ consumption amounts. Some groups have significantly higher consumption than other groups. Therefore, our group is going to select the major group of clients and divide them into a train and test set to conduct a forecast of electricity consumption.

To identify clients belonging to the major group, we need to first remove monthly data in 2011 because there are 210 clients who were created after 2011. Then we decided to drop clients who have at least one month’s consumption between 2012 and 2014 because a zero monthly consumption in either train set or test set would cause a significant bias in our prediction and reduce prediction accuracy. The standard we used to identify the major group is checking whether a client’s monthly consumption from 2012 to 2014 was never above 0.028 \* 10^8kWh since only around 10% of clients have a monthly consumption higher than this value. After processing steps mentioned above, we got 275 clients satisfying major group conditions, which account for approximately 75% of all clients.

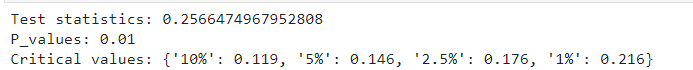
(4) Division of train, validation and test set

We randomly divide 275 clients from major group into train, validation and test set based on two conditions: the ratio between size of (train + validation) set and test set is 4:1, the ratio between size of train set and validation set is 4:1.

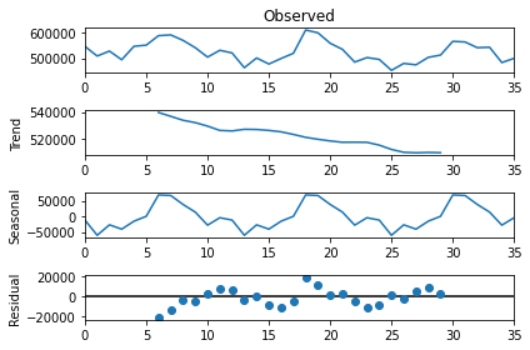
IV. Stationarity Analysis

Since the ARIMA model assumes that the time series is stationary, we should first analyze the stationarity of the time series before applying the model. The test method we chose is the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

The KPSS test has a null hypothesis of “An observed time series is stationary”. For example, we conducted KPSS tests on the train set of monthly electricity consumption for the major group of clients from 2012 to 2014. The result is shown as below:



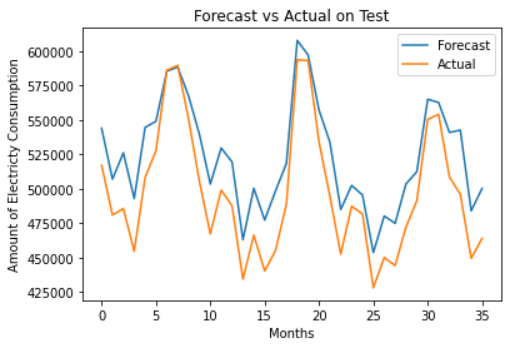
The test statistic is larger than the critical value of 1% significance level and has a p-value lower than 0.01. Therefore, we can reject the null hypothesis and conclude that the train set of major group clients’ monthly electricity consumption is not stationary. We decided to apply the SARIMAX model on this train set based on this result. In fact, there are obvious decreasing trend and seasonality within a 12-month time period from this train set. We can get this information through the “seasonal\_decompose” function from the “statsmodels” package. Here is the plot using this function. The x-label is the number of months from Jan, 2012.



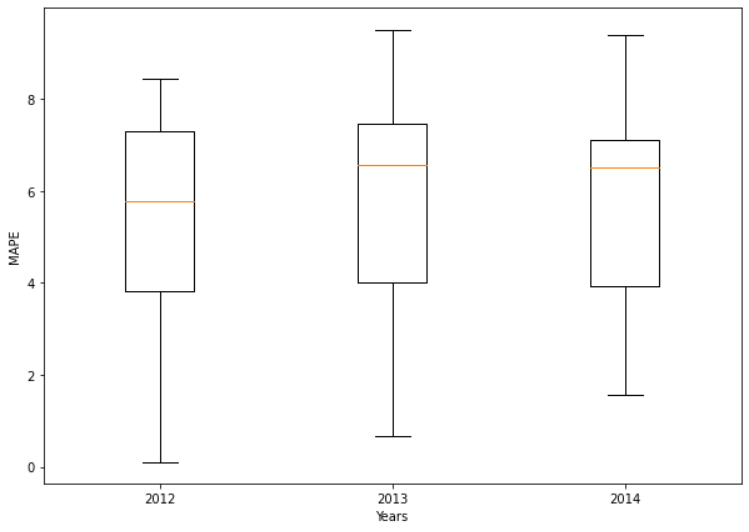
V. Model Training & Prediction Accuracy Analysis

We conducted a hyperparameter tuning process on the train set of clients to find a set of optimal hyperparameter values, which are autoregressive order(p), differencing(d) and moving average order (q). We test the accuracy of different combinations of (p, d, q) while setting length of repeating season order(S) equal to 12 on validation set at first. The Mean Absolute Percentage Error(MAPE) result revealed that the combination of hyperparameters (p = 1, d = 0, q = 0) has the highest accuracy with only 2.29% error rate in the validation set.

Then we constructed the SARIMAX model with these hyperparameter values and applied this model to the test set of clients in predicting their monthly electricity consumption. The comparison plot between predicted monthly consumption for test set and the actual monthly consumption from 2012 Jan to 2014 Dec is shown below:



The overall MAPE result for this model on the test set is 5.60%, which is within a reasonable range. We also split the test set into three equal time periods each with a 12 months length in computing MAPE. Here is a box plot to show MAPE value in each time period in predicting the test set’s monthly consumption.



From the boxplot above we can see the mean MAPE is relatively smaller for 2012, and there’s an approximately increasing trend in MAPE.

VI. Problems Need to be Solved

Since we omitted 95 clients in the process of identifying the major group of clients, we can conduct a monthly electricity consumption analysis on these clients to compare if there is significant difference between their consumption and major group’s consumption. This comparison result can check whether dropping these clients is correct and enhance our accuracy in prediction.

The other problem is that we didn’t include consumption data in 2011 because there were 210 clients who did not have any consumption in this year. Nevertheless, data in 2011 might affect optimal value of trend, seasonality or other parameters, which can lead to accuracy reduction for our model. Thus, if we have time extension, we can construct a model with data in 2011 for prediction and compare its result with our original one.