# ELECTRICITY CONSUMPTION FORECASTING IN PORTUGAL

Ömer Faruk MERCAN, Taylan SABIR, Akif TAŞCI,Emre YILDIZ
Prof. Dr.Murat Can Ganiz

Department of Computer Engineering, Marmara University
Istanbul/Turkey

***Abstract*— In this study we develop forecasting models for electricity demand using K-Means Clustering machine learning algorithm. It compares accuracy of these models using different evaluation metrics. The data consist of 370 different source measurements and observations related to the electricity in Portugal from 2011 to 2014. Our results show that the electricity demand can be forecasted with high accuracy using machine learning algorithms.**

***Keywords*— *Electricity demand forecasting; Time Series Analysis; Machine Learning Algorithms***

## I. INTRODUCTION

Every day, we are surrounded by one of the most important innovations of all time, electricity. While it is a force of energy used all over the world, before discovering it, people have been living for centuries without it, which you could imagine contributed to one dark world at night with the exception of a candle here and there. Nevertheless, even though humans have survived without it, the chances of the human race thriving without it is highly unlikely. Starting with your house, electricity is important for operating all appliances, entertainment, lighting and of course, all technology. When it comes to travelling, electricity is important for the use of electric trains, aeroplanes and even some cars. Consuming electric is important as much as generating it. Most of the countries started to lean on nuclear energy centrals because they can not generate enough electricity. But what if people used electricity sparingness? Would we still needed nuclear energy centrals? In this study, we develop and evaluate prediction models for electricity demand using previous years electricity consumption data.

## II. DATASET

The data subject to our analysis is obtained from Artur Trindade [1].

Data set has no missing values. Values are in kW of each 15 min. To convert values in kWh values must be divided by 4.
Each column represent one client. Some clients were created after 2011. In these cases consumption were considered zero.

All time labels report to Portuguese hour. However all days present 96 measures (24*4). Every year in March time change 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 change day (which has 25 hours) the values between 1:00 am and 2:00 am. The dataset contains 370 different sources elecetricity consumption in kWh with 15 minutes intervals between years 2012 and end of the 2014. These sources could be a house, apartment etc. We don't have a certain information about the source. Our dataset and attribute characterictics is time-series and real. We have 140256 instances and 370 attributes. Data sources does not have any missing values and the type of the attribute is numeric.

TABLE I. Part of the Dataset

| | Date | MT_001 | MT_002 | MT_003 | MT_004 | MT_005 | MT_006 | MT_007 | MT_008 | MT_009 | MT_010 | MT_011 | MT_012 | MT_013 | MT_014 | MT_015 | MT_016 | MT_017 | MT_018 | MT_019 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50993 | 6/15/2012 4:00 | 4 | 21 | 2 | 65 | 28 | 101 | 3 | 175 | 35 | 32 | 23 | 0 | 34 | 24 | 0 | 34 | 30 | 105 | |
| 50994 | 6/15/2012 4:15 | 3 | 21 | 2 | 63 | 28 | 101 | 3 | 165 | 35 | 30 | 24 | 0 | 32 | 22 | 0 | 16 | 31 | 109 | |
| 50995 | 6/15/2012 4:30 | 4 | 21 | 2 | 59 | 26 | 98 | 2 | 168 | 35 | 31 | 26 | 0 | 37 | 23 | 0 | 17 | 29 | 112 | |
| 50996 | 6/15/2012 4:45 | 3 | 20 | 2 | 63 | 24 | 98 | 3 | 152 | 35 | 31 | 24 | 0 | 39 | 23 | 0 | 18 | 31 | 99 | |
| 50997 | 6/15/2012 5:00 | 4 | 20 | 2 | 63 | 24 | 104 | 3 | 168 | 31 | 30 | 25 | 0 | 39 | 22 | 0 | 19 | 29 | 105 | |
| 50998 | 6/15/2012 5:15 | 3 | 21 | 2 | 63 | 26 | 98 | 2 | 155 | 30 | 31 | 23 | 0 | 46 | 24 | 0 | 17 | 29 | 112 | |
| 50999 | 6/15/2012 5:30 | 4 | 21 | 2 | 63 | 30 | 107 | 3 | 172 | 31 | 30 | 22 | 0 | 43 | 26 | 0 | 18 | 28 | 128 | |
| 51000 | 6/15/2012 5:45 | 3 | 21 | 2 | 45 | 18 | 86 | 2 | 199 | 26 | 17 | 19 | 0 | 40 | 19 | 0 | 17 | 29 | 83 | |
| 51001 | 6/15/2012 6:00 | 4 | 21 | 2 | 43 | 20 | 77 | 1 | 148 | 19 | 17 | 17 | 0 | 29 | 19 | 0 | 18 | 28 | 83 | |
| 51002 | 6/15/2012 6:15 | 3 | 22 | 2 | 43 | 21 | 80 | 2 | 178 | 23 | 17 | 18 | 0 | 34 | 21 | 0 | 16 | 25 | 96 | |
| 51003 | 6/15/2012 6:30 | 3 | 8 | 2 | 43 | 18 | 110 | 1 | 158 | 26 | 19 | 21 | 0 | 39 | 23 | 0 | 16 | 26 | 96 | |
| 51004 | 6/15/2012 6:45 | 3 | 9 | 2 | 67 | 18 | 125 | 1 | 155 | 24 | 17 | 20 | 0 | 35 | 24 | 0 | 17 | 28 | 102 | |
| 51005 | 6/15/2012 7:00 | 4 | 12 | 2 | 69 | 20 | 128 | 2 | 182 | 23 | 22 | 25 | 0 | 33 | 21 | 0 | 17 | 33 | 96 | |
| 51006 | 6/15/2012 7:15 | 4 | 22 | 2 | 75 | 20 | 155 | 2 | 182 | 23 | 19 | 23 | 0 | 46 | 22 | 0 | 18 | 41 | 112 | |
| 51007 | 6/15/2012 7:30 | 3 | 23 | 2 | 79 | 23 | 158 | 1 | 182 | 23 | 26 | 28 | 0 | 40 | 25 | 0 | 21 | 40 | 147 | |
| 51008 | 6/15/2012 7:45 | 1 | 24 | 2 | 87 | 27 | 140 | 2 | 192 | 30 | 31 | 28 | 0 | 41 | 33 | 0 | 21 | 39 | 188 | |
| 51009 | 6/15/2012 8:00 | 3 | 21 | 2 | 83 | 23 | 170 | 2 | 175 | 45 | 34 | 28 | 0 | 73 | 37 | 0 | 21 | 41 | 246 | |
| 51010 | 6/15/2012 8:15 | 1 | 21 | 2 | 83 | 26 | 173 | 3 | 192 | 80 | 35 | 28 | 0 | 89 | 42 | 0 | 25 | 48 | 284 | |
| 51011 | 6/15/2012 8:30 | 16 | 21 | 2 | 85 | 27 | 217 | 3 | 242 | 98 | 30 | 34 | 0 | 102 | 47 | 0 | 25 | 41 | 281 | |
| 51012 | 6/15/2012 8:45 | 1 | 22 | 2 | 93 | 24 | 235 | 2 | 232 | 103 | 30 | 34 | 0 | 108 | 56 | 0 | 21 | 39 | 284 | |
| 51013 | 6/15/2012 9:00 | 1 | 18 | 2 | 91 | 22 | 223 | 1 | 269 | 103 | 34 | 40 | 0 | 95 | 62 | 0 | 28 | 47 | 281 | |
| 51014 | 6/15/2012 9:15 | 1 | 20 | 2 | 93 | 26 | 226 | 1 | 263 | 108 | 32 | 40 | 0 | 101 | 63 | 0 | 32 | 43 | 291 | |
| 51015 | 6/15/2012 9:30 | 1 | 18 | 2 | 89 | 24 | 223 | 1 | 300 | 107 | 39 | 46 | 0 | 104 | 64 | 0 | 33 | 44 | 272 | |
| 51016 | 6/15/2012 9:45 | 1 | 20 | 2 | 87 | 28 | 214 | 1 | 293 | 105 | 56 | 44 | 0 | 106 | 60 | 0 | 28 | 49 | 304 | |
| 51017 | 6/15/2012 10:00 | 1 | 18 | 2 | 85 | 34 | 190 | 2 | 303 | 93 | 32 | 36 | 0 | 110 | 63 | 0 | 30 | 46 | 310 | |
| 51018 | 6/15/2012 10:15 | 1 | 21 | 2 | 85 | 43 | 164 | 2 | 313 | 105 | 58 | 39 | 0 | 93 | 56 | 0 | 28 | 43 | 291 | |
| 51019 | 6/15/2012 10:30 | 1 | 23 | 2 | 93 | 39 | 167 | 3 | 296 | 100 | 56 | 37 | 0 | 104 | 64 | 0 | 28 | 46 | 278 | |
| 51020 | 6/15/2012 10:45 | 1 | 18 | 2 | 79 | 38 | 167 | 2 | 303 | 108 | 57 | 39 | 0 | 107 | 61 | 0 | 27 | 54 | 288 | |
| 51021 | 6/15/2012 11:00 | 3 | 16 | 2 | 65 | 43 | 173 | 3 | 296 | 96 | 56 | 39 | 0 | 101 | 61 | 0 | 30 | 53 | 304 | |

**Attiribute Information :** Data set were saved as txt using .csv format, using semi colon (;). First column present date and time as a string with the following format 'yyyy-mm-dd hh:mm:ss'. Other columns present float values with consumption in kW

**Abstarct** : This data set contains electricity consumption of 370 points/clients.

| | | | |
|---|---|---|---|
| Data-Set Characteristics | Time-Series | Number of Instances | 370 |
| Attribute Characteristics | Real | Number of Attributes | 140256 |
| Associated Tasks | Regression,Clustering | Missing values ? | N/A |
| Each Column | Clients | Type of Attribute | Numeric |
| Target(every column) | NO | | |

*(Table 2)*

[1] **artur.trindade '@' elergone.pt, Elergone, NORTE-07-0202-FEDER-038564**

[2] **978-1-5386-0930-9/17/$31.00 ©2017 IEEE**

Our dataset, we don't know exactly our missing value percentage. Because the contributer of our dataset specify that in a table as N/A (not answered). But contributer estimated that missing values are under than 0.5 percentage and greater than 0.1 percentage.

| | | | | | |
|---|---|---|---|---|---|
| Data Set Characteristics: | Time-Series | Number of Instances: | 370 | Area: | Computer |
| Attribute Characteristics: | Real | Number of Attributes: | 140256 | Date Donated | 2015-03-13 |
| Associated Tasks: | Regression, Clustering | Missing Values? | N/A | Number of Web Hits: | 50036 |

Our dataset contains electric consumption of 370 different places.The table shown below is showing some of places' min ,max ,average,standart deviaton and entropy values.

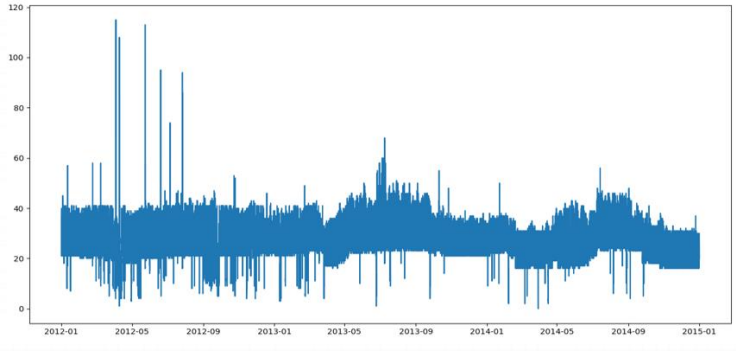
TABLE 2. Min.,Max,Average Values

| Column1 | MİN VALUE | MAX VALUE | AVERAGE | STD. DEVIATION | ENTROPY |
|---|---|---|---|---|---|
| MT_001 | 0 | 48 | 3.98681 | 5.943524 | 19.092938 |
| MT_002 | 0 | 115 | 20.75091 | 13.27231 | 21.472806 |
| MT_003 | 0 | 151 | 3.056696 | 11.00512 | 18.709676 |
| MT_004 | 0 | 321 | 82.15783 | 58.2595 | 23.458029 |
| MT_005 | 0 | 150 | 37.2337 | 26.46517 | 22.31624 |
| MT_006 | 0 | 536 | 141.2385 | 98.43551 | 24.239692 |
| MT_007 | 0 | 45 | 4.519807 | 6.489773 | 19.273964 |
| MT_008 | 0 | 552 | 191.3955 | 121.9764 | 24.678116 |
| MT_009 | 0 | 157 | 39.92053 | 29.80232 | 22.416762 |
| MT_010 | 0 | 199 | 42.19942 | 33.40152 | 22.496854 |

Since there are 370 columns in our data set, we only examined the double correlations between the first 10 columns of our data. The smallest of these correlation values is 0.016861 and the largest one is 0.914424. Correlation values for each column intersect with 1. If the correlation values are close to 1, we have a large number of 0 in our data set. In the binary columns which are close to 1, the numbers 0 and more in the same place, the other values are not very far from each other. The ones with very low correlation values are the ones that are closest to 0; 0 and less in different places, other numerical values have been different from each other.

| | MT_001 | MT_002 | MT_003 | MT_004 | MT_005 | MT_006 | MT_007 | MT_008 | MT_009 | MT_010 |
|---|---|---|---|---|---|---|---|---|---|---|
| MT_001 | 1 | | | | | | | | | |
| MT_002 | 0.390876 | 1 | | | | | | | | |
| MT_003 | 0.141631 | 0.120751 | 1 | | | | | | | |
| MT_004 | 0.300622 | 0.780776 | 0.153734 | 1 | | | | | | |
| MT_005 | 0.296627 | 0.733325 | 0.202995 | 0.914424 | 1 | | | | | |
| MT_006 | 0.31324 | 0.809257 | 0.163676 | 0.93529 | 0.903725 | 1 | | | | |
| MT_007 | 0.166489 | 0.478269 | 0.016861 | 0.392533 | 0.4213 | 0.355628 | 1 | | | |
| MT_008 | 0.360772 | 0.884669 | 0.130662 | 0.899748 | 0.862747 | 0.936157 | 0.432766 | 1 | | |
| MT_009 | 0.354376 | 0.734526 | 0.164674 | 0.798345 | 0.809864 | 0.852617 | 0.328301 | 0.84955 | 1 | |
| MT_010 | 0.228367 | 0.678271 | 0.122634 | 0.725435 | 0.748156 | 0.795832 | 0.316558 | 0.799264 | 0.747657 | 1 |

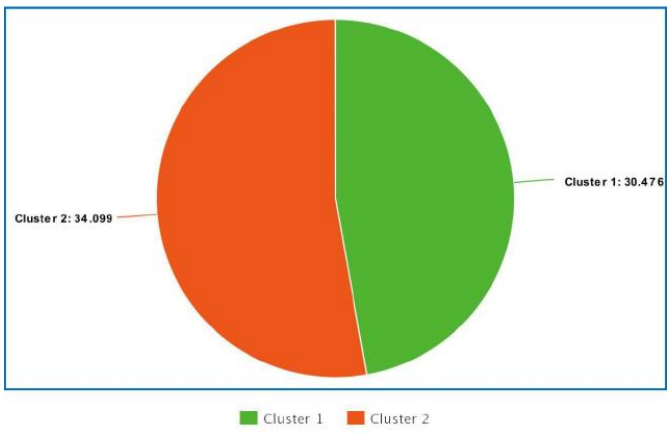
TABLE3. All consumption of a single point for 2011-2014




## III.METHODOLOGY

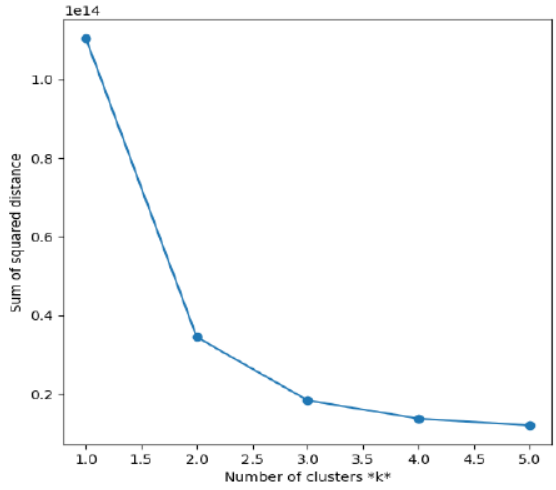
Since our data is time-series data, we initially adapted our data to the k-means algorithm. We divided the time into year, month, day, hour, minute, and we put them in the k-means algorithm along with the data of the electricity usage at 100 points. One of the most important features of the algorithm to choose the 'k' coefficient in a smart way. That's why we applied the elbow method first. Elbow method for the k parameter 2 and 3 selection would be smart, but the most appropriate 2, he said. We also calculated a silhoutte score calculation and the best result would be 'k' parameter 2. We ran our algorithm and plotted the results in a graph. Since the number of columns entered into the algorithm was 105, we could see the result of the algorithm in 105 dimensions. We have just reviewed the day and minute which is the size. Here the cluster centers are two black points. The red dots represent a cluster, while the purple dots represent the other cluster. We tried the K-Means algorithm for different 'k' values and made the score calculations.

TABLE 3. Pie Chart Distribution



For determining to k values, Elbow method is used.As shown in the graph 2 is the most efficinet value for k.

TABLE 4. Elbow Method Results




When looked the Silhoutte-Score and SSE Calculation

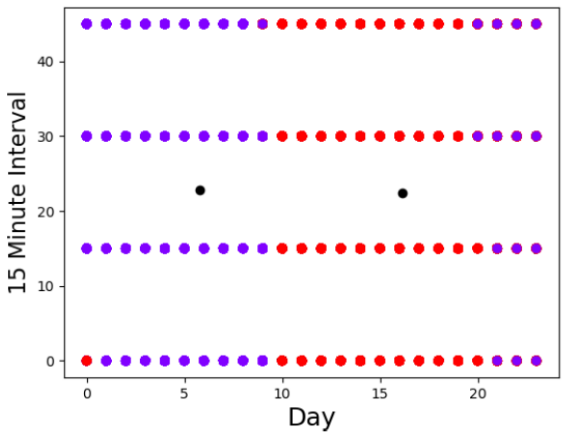
TABLE 5. Silhoutte-Score

- For 2 , 0.616126
- For 3, 0.507959
- For 4, 0.47467

- Error Sum of Squares (SSE) = 65981553677.2014

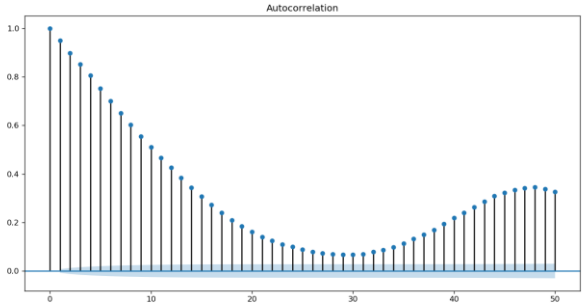
We applied K-means Clustring to our data with k values as 2 and result of algoritm shown in the table below.

TABLE 6. k-Means Algorithm Result




We have not run any feature selection algorithm because our data is time series data. We have an energy expenditure at 370 points. We chose one of these points and we made a prediction with the regression model. The point we selected is MT_002. We used autoregression to predict.

TABLE 7. Autocorrelation




We had three years of data. We used the first two years for the train. We tried to predict the next year. We used an auto regression model for this.

A regression model, such as linear regression, models an output value based on a linear combination of input values.

For example:

yhat = b0 + b1*X1

yhat = b0 + b1*X1

Where yhat is the prediction, b0 and b1 are coefficients found by optimizing the model on training data, and X is an input value. This technique can be used on time series where input variables are taken as observations at previous time steps, called lag variables.

For example, we can predict the value for the next time step (t+1) given the observations at the last two time steps (t-1 and t-2). As a regression model, this would look as follows:

X(t+1) = b0 + b1*X(t-1) + b2*X(t-2)

X(t+1) = b0 + b1*X(t-1) + b2*X(t-2)

An autoregression model is a linear regression model that uses lagged variables as input variables.

We could calculate the linear regression model manually using the LinearRegession class in scikit-learn and manually specify the lag input variables to use.

Alternately, the statsmodels library provides an autoregression model that automatically selects an appropriate lag value using statistical tests and trains a linear regression model. It is provided in the AR class.

We can use this model by first creating the model AR() and then calling fit() to train it on our dataset. This returns an ARResult object.

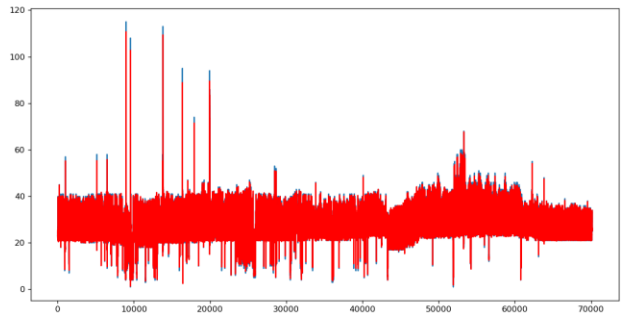
We use an alternative would be to use the learned coefficients and manually make predictions. This requires that the history of 29 prior observations be kept and that the coefficients be retrieved from the model and used in the regression equation to come up with new forecasts.

The coefficients are provided in an array with the intercept term followed by the coefficients for each lag variable starting at t-1 to t-n. We simply need to use them in the right order on the history of observations, as follows:

yhat = b0 + b1*X1 + b2*X2 ... bn*Xn

yhat = b0 + b1*X1 + b2*X2 ... bn*Xn

TABLE 8. Our Prediction Result




An evaluation metric is used to evaluate the effectiveness of information retrieval systems and to justify theoretical and/or pragmatical developments of these systems [10].

Error measurement statistics play a critical role in forecast accuracy.

*R Square*

R square is regression score function. Best possible score is 1.0 and it can be negative. If the r square is 0.97, it means 97 percent accuracy.Our R square value is 0.908

*Mean Absolute Error*

Mean Absolute Error(MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.

MAE = sum(abs(y-y_pred)) / length(y)

Our Mean Absolute Error is 1.254.

*Mean Squared Error*

Mean Squared Error(MSE) is an average of the squares of the difference between the actual observations and those predicted. The squaring of the errors tends to heavily weight statistical outliers, affecting the accuracy of the results.Our Mean Squared Error is 4.402.

## VII. CONCLUSION

In conclusion,by training the first two year ,tried to predict electricity consumpiton of third year with auto regression model which are based machine learning algorithm. $R^2$ (R Square), MAE (Mean Absolute Error) and MSE (Mean Squared Error) evalution metrics were used to test the accuracy of the results. According to the $R^2$, it was observed that up to 90% accuracy was reached with auto regression models.

## REFERENCES

[1] https://towardsdatascience.com

[2] https://www.datascience.com/blog/k-means-clustering

[3] https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

[4] https://blog.cambridgespark.com/how-to-determine the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f

[5] https://scikitlearn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[6] https://www.datacamp.com/community/tutorials/time-series-analysis