Electricity Demand Forecasting based on Smart Meter Data

Gr3at
https://github.com/Gr3at

Abstract

This project aims to study a number of different machine learning regression algorithms in energy consumption time series forecasting using a real world dataset. ARIMA, which is a baseline algorithm for energy forecasting applications, is chosen as benchmark model in this comparative analysis among other well known and widely used machine learning algorithms. Results of this analysis are promising for several studied algorithms, a fact which supports the usage of other algorithms in the field of energy forecasting.

1 Introduction

1.1 Electricity Demand Forecasting

Electricity Demand Forecasting provides valuable information to electricity distribution network operators (EDNOs). EDNOs are county or continent wide companies licensed to distribute electricity to the end user through installed transmission grid. Those companies are also responsible for allocating Meter Points to every consumer of their network, as well as to maintain a database containing energy consumption information for every customer.

1.2 Motivation

Consumption data gathered by EDNOs are very important source of information for Regulatory Authorities (RAs), which are responsible to determine the amount of electrical energy need to be supplied to the network in different periods of time (hour-to-hour, daily), in order to preserve system stability. RAs ideally wish to match customers' energy demands to suppliers' generated power, otherwise various problems arise, which may be in best case scenario wasted energy and in worst case scenario a wide area blackout[1].

A better prediction of electrical energy demand will allow system's operator to schedule power generation in a more efficient manner and as a result the gap between electricity supply and demand will decrease, along with the percentage of overall wasted energy, while concurrently providing stability to the power system.

Stochastic models are designed in order to forecast network demand, based on several features, such as : weather conditions, day of the week, time of the day and customers' profile. The widely used ARIMA model provides very good predictions, but it requires the definition of several hyper parameters tuning in order to extract the best model.

2 Methodology

2.1 Database Description

UK Power Networks[2], an electricity supply company created the dataset in study by collecting data from smart meters installed in several thousands (5.567) of houses in London city. The "London

Households" dataset was downloaded from the official government London Datastore website[3]. It is a file of approximately 10 GB, which contains 167 million rows.

The official dataset consists of historical household power consumption data, along with information about the group each household belongs to (ACORN group) over a period of 3 years approximately. The dataset contains energy consumption data, in kWh (per half hour), unique household identifier, date and time and ACORN Group[4]. ACORN Group classifies consumers to several groups based on many different criteria, such as : ethnicity, age, economic and access to new technologies.

Since, electricity consumption is strongly dependent to weather conditions, an additional dataset containing weather related data had to be used. Weather data were collected using Darksky API[5], which provides access to many different weather conditions data (humidity, temperature, wind, etc) for the city of London for the same period of time. Combining all above features in a single dataset will hopefully help in building a better predictive model.

Unfortunately the provided data are sparse for each household and thous several preprocessing steps need to be followed.

2.2 Preprocessing

During the preprocessing procedure both electricity and weather datasets were modified before their final combination.

For the electricity consumption dataset, the following steps took place. Firstly, the original 167 million rows dataset was split into smaller manageable datasets. Each sub-dataset was modified from string input values to numerical or categorical values and DateTime cell were defined as such. Nan values were replaced with min(value) of their associated categories. Since we were interested to study overall electricity consumption, columns containing household specific information are of no use and thous those columns were dropped. After that data were resampled to 1 hour frequency and groups by ACORN Groups were created, resulting in great dataset size reduction, from 167 million rows (sum of all sub-datasets) to 735 thousand rows in total.

Regarding the weather dataset, everything was much easier. Weather data were pretty much clean (no Nan values, good sampling frequency of 30 minutes). The only preprocessing applied in this case is filtering for odd values in the data (temperature[-20 to 40 degrees], wind speed[0 to 150 km/h] and humidity [0% to 100%]) and several unused features were removed.

Once both datasets were clean and equally indexed (same time indexing), they were combined into one dataset. Data were then normalized and scaled, so that machine learning algorithms could be applied on them. Since our dataset is time dependent, we need previous lags of our time series to predict current or future lags. Thous further preprocessing needs to be applied, in order to create new features representing at least t-1 state of our features. In order to decide how many previous states are helpful for our predictions, autocorrelation of the feature of interest has to be implemented. As you can see in Figure 1 lower right sub-figure, first 3 timesteps of the feature of interest are highly correlated.

A visual observation of electricity demand data would be very informative and would help to better understand the dataset. Having that in mind, a graph of the total energy demand (sum energy of all ACORN Groups) over time is provided in Figure 1 upper sub-figure.

It is obvious that energy consumption before June 2012 has suspiciously very low values and thous we assume that data collected before this date are not accepted and will be removed before our analysis takes place.

Before getting into machine learning we should also plot Energy consumption by Acorn Group over time. There are 6 different Acorn Groups, as analyzed in [4], each of which contains a wide group of households based on several characteristics. Figure 1 lower left sub-figure shows which Acorn Groups are the most energy consuming and which have the lowest fingerprint.

Figure 2 provides information regarding the dataset. Specifically, we can see how features of interest (electricity demand and temperature) are distributed, with the help of histograms. Moreover a representation of the hourly distribution of electricity demand over a period of one year and finally a heatmap (blue to red represent low to high daily consumption) showing how electricity demand changes regarding month of the year and day of the week are provided in the same figure.

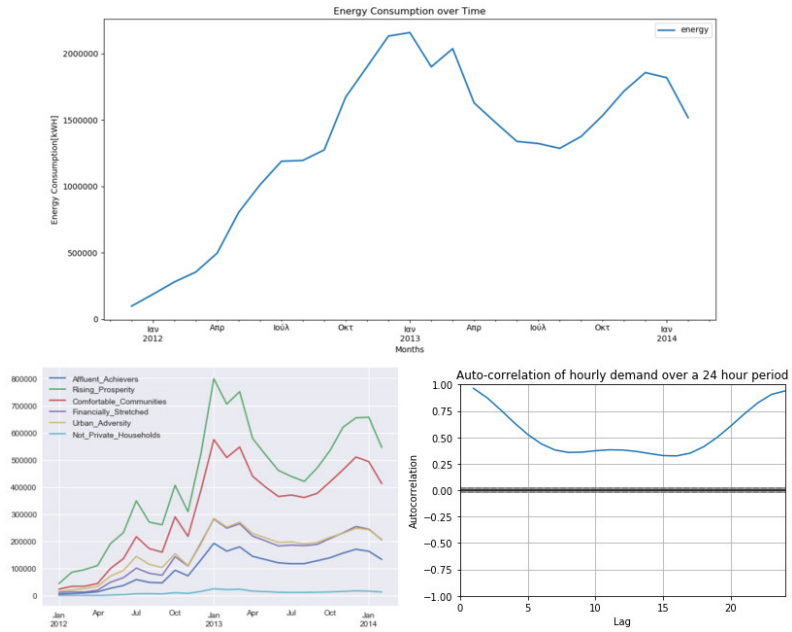

Figure 1: Electricity Demand Exploration.

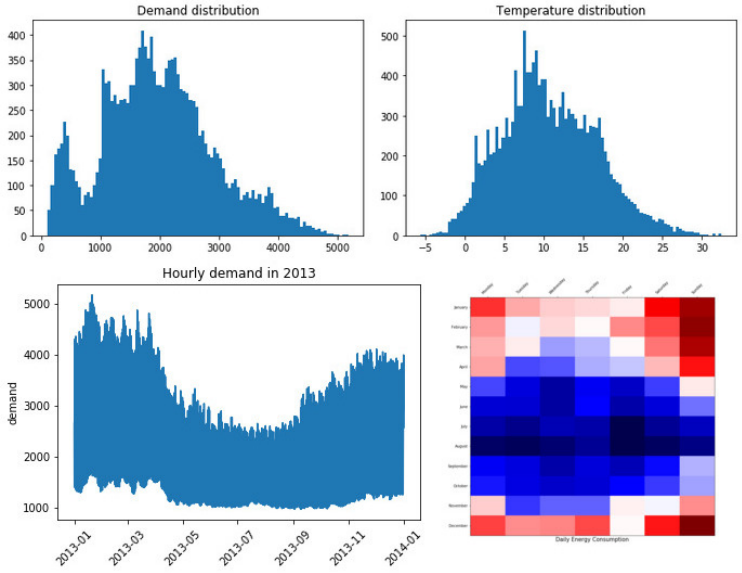

Figure 2: Exploratory Analysis of the Data

2.3 Approach Followed

In order to study the stated problem and define new approaches in Electricity Demand Forecasting, several Regression Algorithms were applied to our dataset. The studied algorithms are : Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector (SVR) and Multi-Layer Perceptron (MLP) Regression Models. Analysis of the dataset, preprocessing and all above mentioned models were implemented fully in python, using Spyder ide and Jupyter notebook. For the process of model selection for each algorithm, a 5 fold cross-validation was chosen.

2.3.1 Autoregressive integrated moving average - ARIMA

ARIMA models are the most widely used in the field of energy demand forecasting. We will use a Seasonal ARIMA model as the baseline benchmark model. For our S-ARIMA [7][8] best model selection, only order (selected-(2,1,0)) and seasonal-order (selected-(1,1,0,24)) parameters were tuned.

2.3.2 Linear Regression

LR was the first tested model. Its advantage of few tuning parameters and fast execution time, were the reasons it was selected to extract first exploratory results and understand which features have the greatest impact in electricity demand forecasting.

2.3.3 Decision Tree and Random Forest Regression

Both algorithms are very fast and provide satisfactory results in most of the cases. In RF model we tuned : #trees, max tree depth and #features used in each tree and the best model return was RF [6] (500 trees, 'auto' #features, 14 max tree depth).

Our DT model was tested for a combination of different splitting criteria, maximum depth and #features used and the returned model of Cross-Validation was DT('mae','auto',11).

2.3.4 Support Vector Regression - SVR

Support Vector Machines [6] are among the state of the art algorithms used in machine learning and thous we were curious to see how well they would adopt our problem. The model was tuned in regards to the kernel function used (tested : poly, rbf and sigmoid), the cost 'C' applied to outlier and gamma coefficient of each kernel and the selected model was SVR('rbf',0.072, 7).

2.3.5 Multi-Layer Perceptron - MLP

Models based on MLPs [6] are among the most complex, easy to overfit, but appear to have great performance. There are a lot of parameters to tune in this case, but the ones selected are: the activation function (identity, logistic, tanh, relu), network structure (# hidden layers and #perceptrons in each layer) and solver for weight optimization (lbfgs, sgd, adam). The selected model was MLP(relu, (24,7,52), 'lbfgs').

We didn't use brute force approach for #hidden layer and #perceptrons selection, rather we used specific combination based on intuition. The one selected for example has 3 hidden layers. The first layer has 24 perceptrons, same as hours in a day, the second layer has 7, same as days in a week and the third has 52, just like weeks of a year.

2.4 Evaluation Metrics

Our selected models were evaluated based on their test Root Mean Squared Error (RMSE), R-Squared score and Total model train-test execution time. RMSE provides information on the similarity between predicted and actual values. As predictions and actual values limit to equality, value of RMSE approaches zero. R-Squared score explains how well does a model fit data of interest. Eventually, a value close to 1 implies that the model fits provided data very well. Total execution time is also an important metric, since our dataset consist of hourly electricity demand predictions and if it is to be used in the field, it has to fulfill time requirements.

3 Experiments - Analysis Results

As mentioned in section 1, the objective of this analysis is to compare various predictive models regarding electricity demand forecasting.

After following preprocessing steps described in section 2.2, we trained each of the models described in section 2.3 using a number of different configurations and we extracted the best model of each algorithm, based on specific 5-fold cross-validation. A representation of the cross-validation approach

is given in figure 3. It is important to respect fold sequence, because intuitively it wouldn't hold to predict previous values from future values.

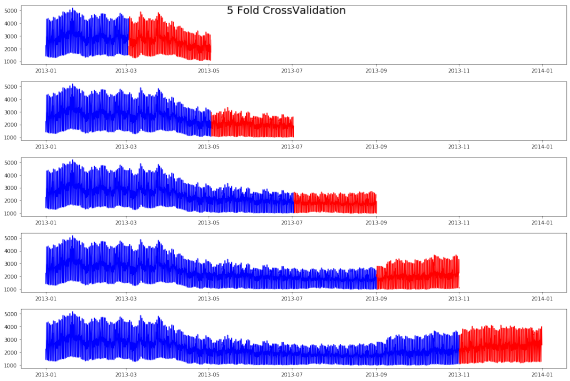

Figure 3: Cross-Validation Approach Used for the Study

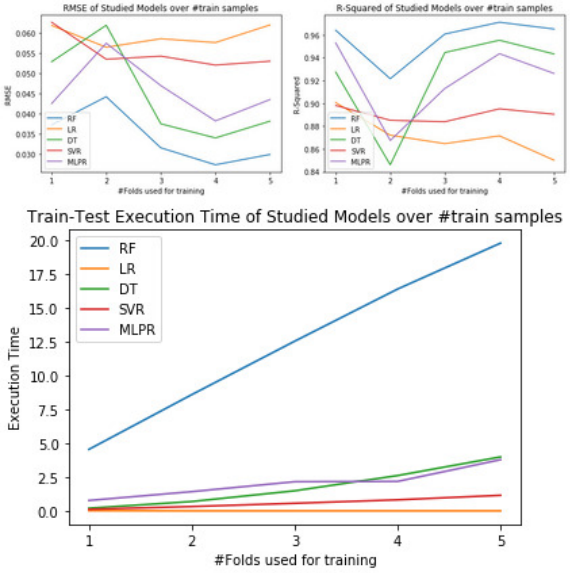

Figure 4: Cross-Validation Evaluation Metrics

Figure 4 shows evaluation metrics values for all best models by the process of adding more folds to the training phase. An interesting observation is that all models have lower performance in the second fold than in the first one and that is maybe due to the fact of the increasing number of unseen validation samples. That is why from the next fold metrics seem to improve in performance.

The next step in our analysis consists of training the set of best selected models using all 5 folds. Once the models are trained we initiate the comparative analysis by testing them using a test set, containing roughly 2 months of data, starting at 2014-01-01 and ending at 2014-02-28. Table 1 shows results of all models after testing with the test set.

Some of the results were not expected. Random Forest is a rather fast and accurate algorithm. RMSE and R-Squared scores confirm our knowledge, since from both perspectives RF is the best model. Total model execution time on the other hand was too long. Of course the number of trees (500 in this case) is a mandatory parameter to the total execution time, but we did not expect the difference from the other models to be so big.

Regarding SARIMA model, results are discouraging. RMS error is not that bad, when compared to other models, but R-Squared score is very bad. A negative value implies that SARIMA model

Table 1: Model Performance on Test Set

Model	RMSE	R-Squared	Elapsed Time
SARIMA	0.088	-0.076	8.324
RF	0.024	0.977	65.098
LR	0.036	0.946	0.010
DT	0.036	0.947	8.920
SVM	0.062	0.840	0.254
MLP	0.033	0.954	2.672

doesn't fit test data. Results suggest that more inspection is required in model parameters, in order to get better results.

4 Conclusion

In this study several different machine learning algorithms were implemented and tested for a case study of electricity demand forecasting. Results are very promising for most of the tested algorithms and thous more investigation of these models and their performance using similar dataset of the field of study would be a good step to strengthen their reputation before electricity aggregators use them in real world applications.

Figure 5 provides a graphical representation of the actual versus the predicted next hour electricity demand forecast in the upper sub-figure as well as a 3-hours forecast error in the lower sub-figure. Predicted values are outcomes of the Multi-Layer Perceptron trained in our study.

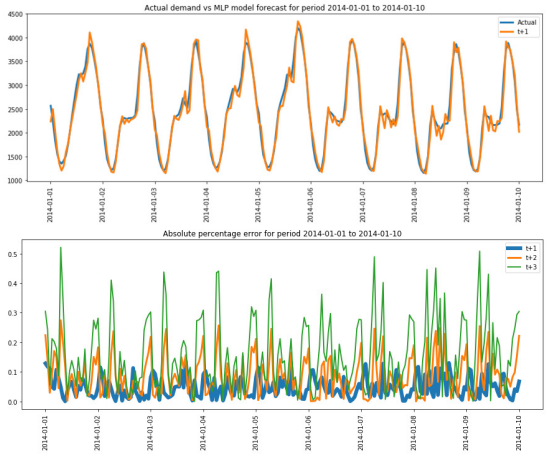

Figure 5: MLP Electricity Demand Forecasting and Forecasting Error

It would be interesting to repeat parts of the analysis made, in future studies where electricity demand forecasting of different frequency could take place. An idea would be to create predictive models for daily electricity demand. A model of this type could be very useful for next day planning. Moreover a weekly or even a monthly basis forecasting would help Power Plants Engineers to decide the best possible time to take out generators for maintenance.

Acknowledgements

The studied dataset, found at Kaggle (link), was collected by Jean-Michel Daignan.

References

[1] Marko Cepin & Radim Bris (2017) Safety and Reliability. Theory and Applications, CRC Press

[2] UK Power Networks. www.ukpowernetworks.co.uk

[3] London DataStore. data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[4] Acorn. acorn.caci.co.uk/downloads/Acorn-User-guide.pdf

[5] DarkSky API. darksky.net/dev

[6] scikit-learn Documentation. scikit-learn.org/stable/supervised_learning.html

[7] statsmodels documentation. www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

[8] Wei-Chiang Hong (2013) Intelligent Energy Demand Forecasting, Springer