SMART METER ANALYSIS

Karthik Ram

Karthikram398@gmail.com

Abstract

The ability to predict consumption is an essential tool for the management of a power distribution network. The availability of an advanced metering infrastructure through smart meters makes it possible to produce consumption forecasts down to the level of the individual user and to introduce intelligence and control at every level of the grid. While aggregate load forecasting is a mature technology, single-user forecasting is a more difficult problem to address due to the multiple factors affecting consumption, which are not always easily predictable. This work presents a hybrid machine learning methodology based on random forest (RF) and linear regression (LR) for the deterministic and probabilistic forecast of household consumption at different time horizons and resolutions. The approach is based on the separation of long-term effects (RF) from short term ones (LR), producing deterministic and probabilistic forecasts. The proposed procedure is applied to a public dataset, achieving a deterministic forecast accuracy much higher than other methodologies, in all scenarios analyzed. This covers horizons of the forecast from one minute to one year and highlights the great added value provided by probabilistic forecasting.

Keywords: Random forest, Linear regression, SVM, ANN, Hybrid machine learning, CRBM, FCRBM

1.Introduction

Energy expenditures will be lowered by increasing the possibility of reduced consumption using analyzed Smart Meter data motivated to perform this research work. During the usage of traditional meters, there is the involvement of wastage of much energy to manpower. As the electricity consumption of the household is known on monthly basis by conventional meters, there is an overall demand for the electricity utilities to explore a new development for benefit of the consumers as well as themselves. However, the study determines to make attempts to replace electricity meter in respective households by minimizing the drawbacks that occurred by the consumer. The daily electrical usages change concerning habits and it is mostly dependent on the behavior of consumers. By using traditional meters, usages are not flattened as consumers are not aware of the knowledge about how much consumption has been made in an hour or any interval of time in a day. The uncertain perception of the consumers can also be falsified as most of the consumers have very low knowledge regarding the Smart Meter and its installation. Lastly, to enable change and read concerns in the market also motivated to perform this study.

1.2 Problem Statement:

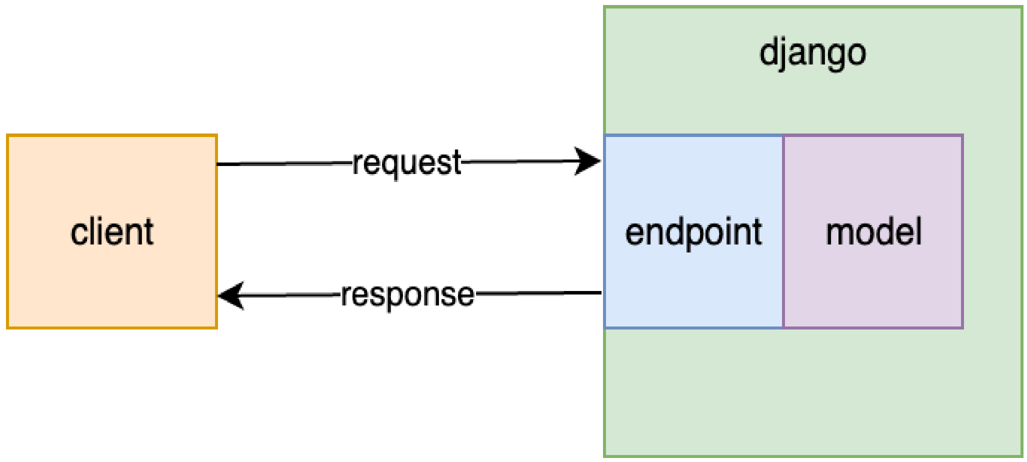
How to use the volume of data from smart meters to promote and improve efficiency and sustainability of demand has become a major research topic worldwide. Control decisions for the smart grid should be made continuously at both the aggregate and granular levels. Power distribution companies rely on forecasts with different time horizons to support both system operability and planning. Retail electricity suppliers are making pricing, procurement and hedging decisions based largely on the expected load of their customers greater use of renewable energy sources, which results in a load profile increasingly characterized by the presence of peaks in consumption due to human behavior, that can lead to problems for electricity providers. This work has found application the forecast of energy consumption over time allows property and building managers to plan energy consumption over time, shifting energy use to off-peak periods, improving energy purchase plans, and allowing them to assess their consumption habits.

2. Literature review:

To address the problem of load forecasting at the level of smart meters, the research community has attempted different approaches, from adapting techniques already widely used for aggregate load for recasting, to developing new techniques or using a combination of the same. Methods such as the semi-parametric additive model, exponential smoothing, and classical seasonal time series methods have been applied to load forecasts at the building level, as well as methods based on artificial neural networks(ANN)and support vector machines(SVM). A 2012 study compared several existing techniques, including linear regression (LR) as well as different types of ANNs and SVMs on two datasets: one for two commercial buildings and the other for three residential homes. The results showed that the techniques used could provide reliable forecasts in the first case but not in the second, because of the greater variability of the load. In the load, the forecast is studied both at the building level and the state and provincial level through a self-recurrent wavelet neural network. The recent trend is to use deep learning techniques, with recursive or convolutional or hybrid neural networks. The conditional restricted Boltzmann machine (CRBM) and factored conditional restricted Boltzmann machine (FCRBM) have been evaluated to estimate the energy consumption of a household. The FCRBM achieves the highest accuracy in load forecasting compared to ANN, RNN (Recurrent Neural Network), SVM, and CRBM. Different resolutions ranging from one minute to one week have been tested. The same dataset has been analysed in two successive works in which the effectiveness of recursive networks of the type long short term memory (LSTM) has been investigated both in the standard form and in the form sequence to sequence (S2S) and in which the accuracy of convolutional neural networks (CNNs) has been evaluated, obtaining comparable or superior results to those obtained with the FCRBM algorithm. Due to the high variability of smart meter measurements, it may be necessary to perform probabilistic forecasting in operational practice. Probabilistic load forecasting has also been carried out on individual load profiles; in a method combining gradient boosting (GB) and quantile regression has been proposed to quantify uncertainty and generate probabilistic forecasts, the conditional kernel density(CKD) method have been tested. Recently, a point and probabilistic forecast of the load for 100 low voltage (LV) feeders has been conducted in comparing several methods such as Holt-Winters-Taylor seasonal exponential smoothing, kernel density estimation, seasonal linear regression, and two autoregressive (AR) methods.

3. Methodology and framework:

3.1 System architecture:



3.2 Equations and the proposed method

In the training phase, the technique of bootstrap aggregating (bagging) is applied to tree learning, through which a random subset with replacement from the training set is selected Btimes, the subset of samples from the training set X is identified with Xb ⊂ X and the corresponding label Yb ⊂Y. For each of these subsets, a tree FB is fitted. In the decision trees training process additional randomization, called feature bagging, is used, which consists of considering for each candidate split a random subset of features. After training, the prediction for an x sample is obtained by averaging the predictions of all the generated regression trees: ˆ yet = B ∑ b=1 fb(x) (2) Our procedure uses a forest of 100 trees, whose optimal depth is established through a cross-validation procedure. The features used are only of a temporal type and uniquely identify each measurement: the year, the day of the year, the day of the week, and the time of the day expressed as a real number including fractions of an hour. Many of the recently proposed algorithms for predicting household consumption use a purely auto-regressive approach using non-linear methods based on neural networks. For the short term component of the forecast, we will use an approach based on a simple and fast step-wise multiple linear regression (MLR). It is given by:

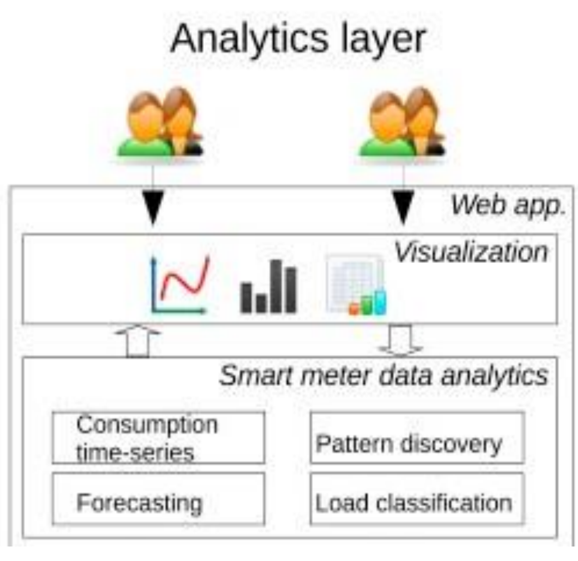
$$\hat{y}_{st}(t + j\Delta t) = \beta_{0,j} + \sum_{i=1}^{n_i} \beta_{i,j} r(t - (j-1)\Delta t),$$

with

$$r(t) = y(t) - \hat{y}_{lt}(t).$$

The short-term forecast for each time t+ jΔt, with j = 1,..., L, is a linear combination of the residual r of the long-term forecast relative to the actual value of power consumption for the most recent ni steps. βi,j denotes the regression parameters. The forecast is obtained simultaneously for NJ's future steps. Given the great variance of domestic consumption, it is important to provide an estimate of the forecast error for each forecast time. This can be obtained simply by analyzing the distribution of the error that the model has made in the training set and applying the same distribution of the error also in the forecasting phase. The distribution of the error is described by 19 quantiles (q = 0.05,0.10,...,0.95), and is parameterized according to the time of forecast j and for each hour of the day h, thus creating a look-up table Eq(j,h) to be applied even in the forecast phase. The procedures described were implemented using only public domain tools. The code is written entirely in python, using the pandas libraries sci-kit-learn for machine learning tools. All libraries have been used in the latest revision available at the time of writing. The procedure has minimal requirements for its execution, the time required for training and forecasting are listed in Table 3, a PC with 16 GB of RAM and a quad-core Intel i5 processor at 3.2 GHz has been used for the analysis.

4. Work done:



4.1. Results:

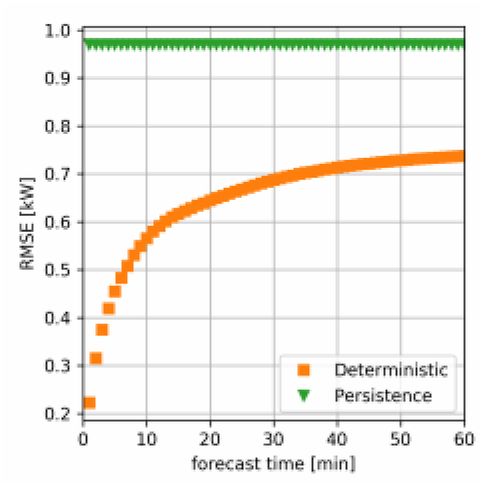
Some load forecast scenarios are analysed, based on the historical power series, differentiated based on the sampling frequency used and the forecast horizon. The scenarios analysed are shown in Table. The selection of the scenarios has been carried out also according to the existing literature on the same dataset to allow a direct comparison of the accuracy of the proposed method.

Forecast with One Minute Resolution:

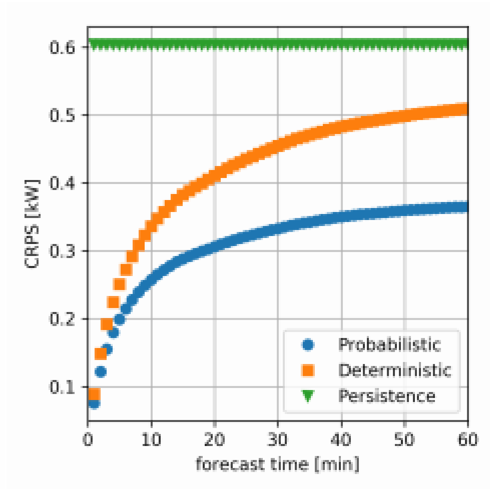
The scenario provides for a time resolution of one minute with forecast times of up to one hour. This resolution was chosen because this time scale is used in the operation of the utility system and real-time market activities, for example in automatic control of generation and resource redirection. Such a high resolution can also be used by home automation systems to prevent the risk of power cut-off due to overloading, as well as to operate small storage buffer systems. The autoregressive short-term component of the forecast is based on the measurements of the hour before the time of forecast. The number of lookback steps is, therefore, equal to 60, the same as the forecast steps. The below table represents the results of the forecasting procedure for the various scenarios analysed. The optimal depth of the tree for the Random Forest algorithm used for long term forecasting, the values of the deterministic errors, mean absolute error(MAE) and root mean square error(RMSE), the skill score(SS)for the RMSE concerning persistence, the accuracy of the probabilistic forecast measured with the continuously ranked probability score (CRPS), and the computational time required for the algorithm's training (including the search for the optimal depth of the trees) and the emission of a forecast are listed.

Forecast with One Week Resolution:

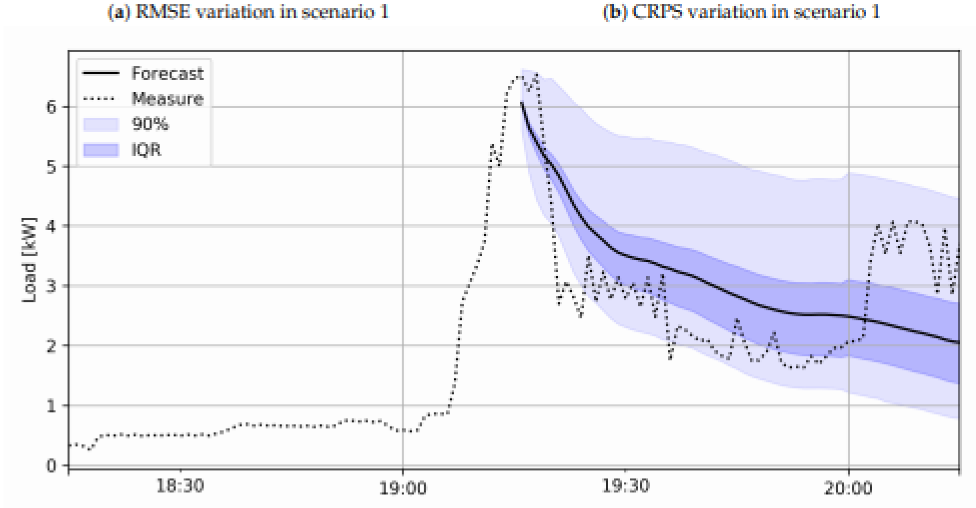
Scenario with a one-week resolution for a forecast with a very long horizon of one year. In addition to assessing the accuracy of the proposed methodology and its applicability to a wide range of time resolutions and forecast horizons, this scenario can be used in the household to verify the trend of its consumption and assess the effectiveness of any changes in habits that may lead to the more rational use of energy resources. In this case, the short-term component of the forecast was not used, and the forecast with persistence used the measurement of consumption one year before the expected time. Both the fitting obtained with the model and the forecast for the final year compared with the aggregate week measurements of the dataset has been represented



(a) RMSE variation in scenario 1

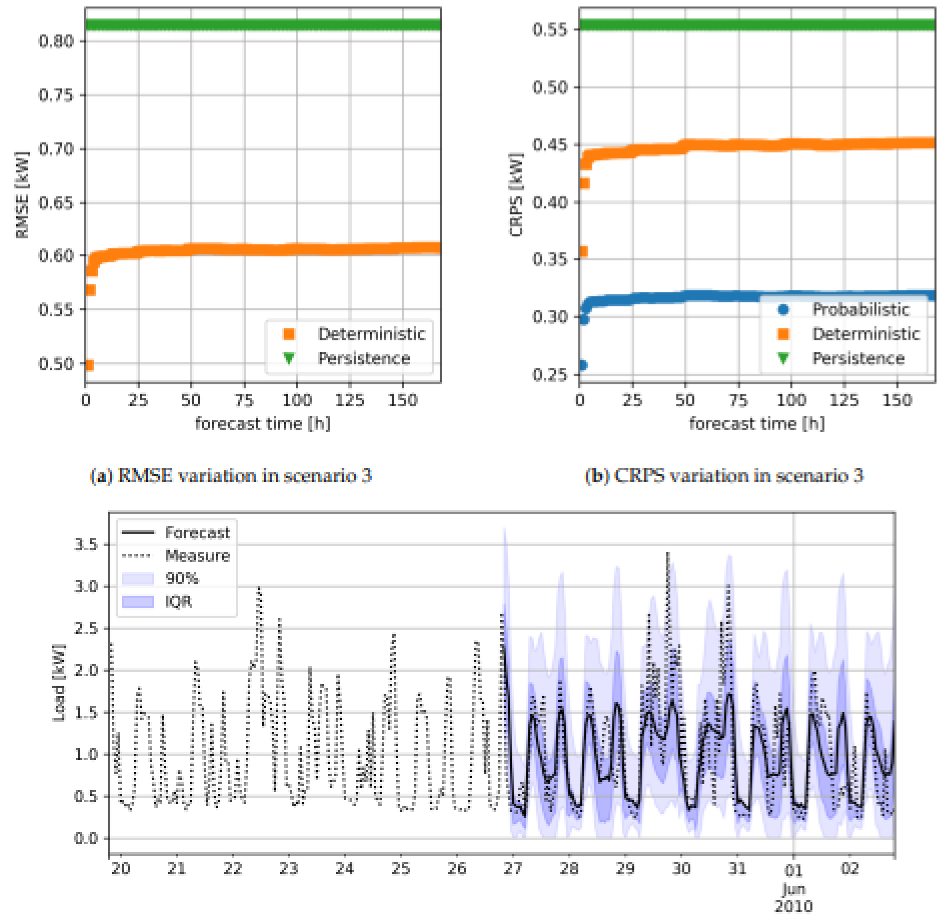


(b) CRPS variation in scenario 1



(c) Forecast for scenario 1

Results for scenario 1 (forecast interval 1 h, sampling rate 1 min): (a) Variation of the RMSE as a function of the forecast horizon, 60 steps correspond to 1 hr; (b) variation of the probabilistic and deterministic CRPS as a function of the forecast horizon; (c) measured load, consumption forecast and probabilistic prediction intervals with the forecast issued at 19:15 on 26 May 2010



(a) RMSE variation in scenario 3

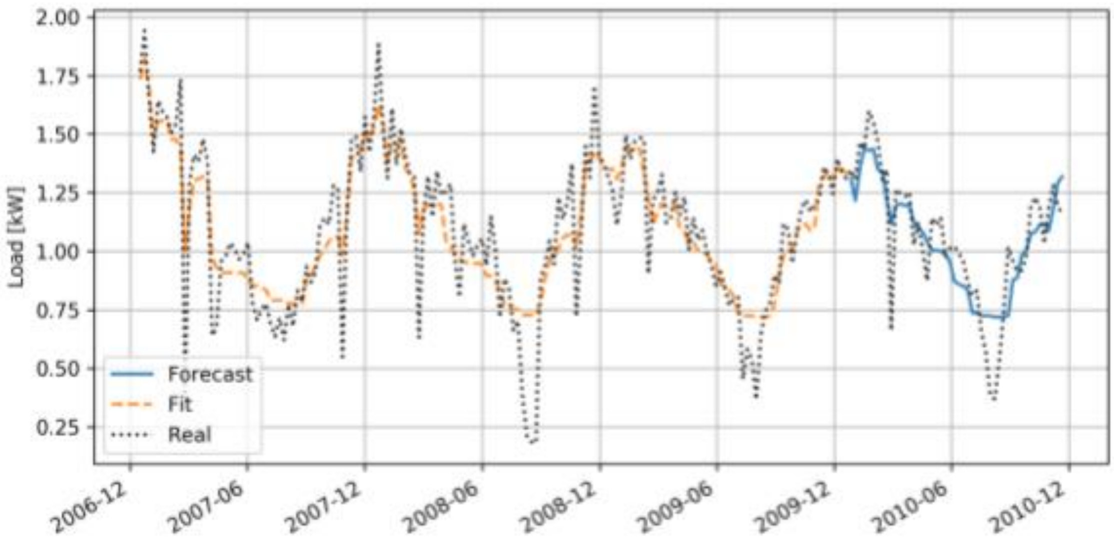
(b) CRPS variation in scenario 3

(c) Forecast for scenario 3

Results for scenario 2(forecast interval 1 week, sampling rate 1 h): (a) Variation of the RMSE as a function of the forecast horizon, 168 steps correspond to 7 days; (b) variation of the probabilistic and deterministic CRPS as a function of the forecast horizon; (c) measured load, consumption forecast and probabilistic prediction intervals with the forecast issued at 19:15(26th May)

Scenario	Description	Method	RMSE (kW)
1	1 h forecast, 1 min resolution	ANN	0.732
		SVM	1.995
		RNN	0.939
		CRBM	0.903
		FCRBM	0.666
		S2S LSTM	0.667
		RF + LR	**0.648**
2	1 day forecast, 15 min resolution	ANN	0.907
		SVM	1.344
		RNN	1.009
		CRBM	1.030
		FCRBM	0.899
		RF + LR	**0.704**
3	1 week forecast, 1 h resolution	ANN	0.785
		SVM	0.791
		RNN	0.916
		CRBM	0.691
		FCRBM	0.663
		S2S LSTM	0.625
		CNN	0.677
		RF + LR	**0.604**
4	1 year forecast, 1 week resolution	ANN	0.246
		SVM	0.188
		RNN	0.457
		CRBM	0.182
		FCRBM	0.170
		RF + LR	**0.145**

The above table represents a comparison of RMSE with the results of other forecasting procedures available in the literature. Artificial neural network (ANN), support vector machine (SVM), recurrent neural network (RNN), conditional restricted Boltzmann machine (CRBM) and factored conditional restricted Boltzmann machine (FCRBM) results; sequence to sequence long short term memory (S2S LSTM) and convolutional neural network (CNN). Random forest (RF) and linear regression (LR)



. The figure shows results for scenario (average weekly consumption from 1 January 2010). The plot shows the measurements available in the dataset, the fit of the model in the training set consisting of the first 3 years of measurements, and the forecast obtained for the last year.

4.2. Discussion:

The public availability of the examined data set allows an indirect comparison of the performances obtained with the proposed methodology with previous works that have analyzed the same dataset. The selection of the scenarios was partially linked to the possibility of making such a comparison. The table shows the values obtained in the literature using machine learning methodologies and deep learning techniques with recursive and convolutional neural networks already cited in the introduction. The considerably better result obtained with the proposed methodology is certainly linked to the choice of a hybrid approach in which the forecasting of phenomena in the long term, characterized by more seasonality (daily, weekly and annual) were separated from the forecasting of the load in the short term, which was instead estimated with a simple autoregressive methodology. The better performances can be only partially attributed to the chosen regression algorithm, it would probably be possible to obtain similar performances using neural networks suitably trained both for the long term component and for the short term component of our model, at the price of a more complicated feature pre-processing and of longer calculation times for the training. The use of a hybrid methodology allows instead to effectively isolate the average behavior from the alterations of the same and to train a regression model more quickly and effectively.

5. Acknowledgments :

This work aims to implement a forecasting procedure for household consumption using only smart-meter load data, with different time horizons and producing both deterministic and probabilistic forecasts. The most recent research sees increased use of deep learning techniques for load forecasting, their real effectiveness for forecasting time series with seasonality is however at least partially questioned based on the results obtained with these methods when compared with classical statistical methods. S.Makridakis in a recent article analyses the performance of the methods proposed for M3 competition, highlighting how statistical methods allow obtaining, on average, more accurate forecasts and with a lower computational cost. It also provides suggestions on how to exploit the undoubted potential of machine learning (ML) techniques also in the context of time series forecasting. In particular, it highlights, among other things, the need for data pretreatment, applying transformations and procedures of de-trending and de-seasonalization that allow obtaining a stationary signal, and the accurate assessment of the risk of overfitting for ML procedures. In our case, the indications translation to a separation between the long-term components of the signal, associated with the trend and seasonality, and the short-term components linked to stochastic variations in the average trend of consumption of a consumer, which we will treat as two distinct problems in a hybrid methodology. The method we propose provides for an estimate of the average long-term behaviour of users, to which a short-term auto-regressive forecast is superimposed, linked to the deviation of the previous estimate from the most recent consumption measures. A similar hybrid approach, in which the effects of long and short term are separated, has been recently used by S. Smyl in the winning procedure of the recent M4 competition. The proposed methodology is implemented using ML techniques with limited computational requirements to allow an implementation also in low-cost devices installed directly at the user's household. We will use the random forest (RF) technique for long-term forecasting and a simple linear regression for short-term forecasting. We will also realize a probabilistic forecast through a simple persistence of the distribution of forecast errors measured in the training dataset. This work will highlight

• The convenience of a hybrid approach, which separates long-term and shortterm effects for load forecasting when using machine learning techniques.

• The effectiveness of the proposed procedure for predicting smart meter loads.

• The relative contribution of these two components to the accuracy of forecasting.

• The importance and added value of a probabilistic forecast for household load prediction

6. Conclusion:

This work presents a hybrid machine learning methodology based on random forest and linear regression for the deterministic and probabilistic prediction of household consumption at different time horizons and resolutions. The approach is based on the combined forecasting of long and short periods, using in the first case temporal features for the identification of trends and various seasonality's of the time series, and, in the second case, an auto-regressive approach using the most recent load measurements available at the time of emission of the forecast. Finally, through the analysis of the forecast error of the model, a probabilistic load forecast is realized. The analysis highlights the relationship between the accuracy of the forecast and the lead time of the forecast and the time of the forecast and compares the result obtained both with a reference based on persistence and with the results of the analysis of the same dataset with machine learning and deep learning methodologies published in the literature. The method proves to be very effective in terms of absolute precision and calculation times required for model training and forecasting. The method also offers opportunities for development for both the long term and short-term components, using additional predictors, experimenting with non-linear methods for short-term forecasting and refining the probabilistic forecasting methodology

6.1 Future scope:

The proposed methodology can naturally be improved both for the long term and the short-term part. For the long term, we are planning to introduce predictors based on the weather conditions and holidays scheduled in the calendar. For the short term, it may be interesting to evaluate the use of a non-linear regressor, which could be based on RF as well. For both components, additional features can be represented by the sub-meter readings provided as additional information in the dataset. We expect that a possible improvement can also come from the adoption of pattern recognition techniques for the identification of the activation of the appliance, which can allow a higher precision in identifying consumption habits by the user. • Develop the most reliable website for all kinds of constraints exhibited by the user at households getting maximum accuracy with the model.

7. References:

1.Farhangi, H. The path of the smart Grid. IEEE Power Energy Mag.2010

2.Mohassel, R.R.; Fung, A.; Mohammadi, F.; Raahemifar, K. A survey on advanced metering infrastructure. J. Electr. Power Energy Syst.2014,63, 473–484.

3.Wang, Y.; Chen, Q.; Hong, T.; Kang, C. Review of smart meter data 4. Smart Meter Forecasting from One Minute to One Year Horizons

4. Gajowniczek, K.; Nafkha, R.; Za˛bkowski, T. Electricity peak demand classification with artificial neural networks. In Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 3–6 September 2017; pp. 307–315 . 5. Massidda, L.; Marrocu, M. Decoupling Weather Influence from User Habits for an Optimal Electric Load Forecast System. Energies 2017, 10, 2171.

6. Mocanu, E.; Nguyen, P.H.; Gibescu, M.; Kling, W.L. Deep learning for estimating building energy consumption. Sustain. Energy Grids Netw. 2016, 6, 91–99.

7. Yu, C.N.; Mirowski, P.; Ho, T.K. A sparse coding approach to household electricity demand forecasting in smart grids. IEEE Trans. Smart Grid 2017, 8, 738–748.