Forecasting Electricity Consumption Using Simulation Data from Physical Models

Abdul Rehman
abdul.rehman@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Portugal
December 2019

Keywords: Load forecasting, Random Forest, Energy Services

ABSTRACT – Ever growing population has given rise to the electricity demand which results in more Greenhouse gas emissions connected to its generation and consumption. Currently buildings accounts for the 30 % of CO_2 emissions globally. These alarming numbers have given rise to concepts of the building energy management systems (BEMS) and smart grids. BEMS has been the topic of research since last four decades using numerical techniques but the recent developments in the use of machine learning (ML) technologies for load forecasting has shown a great potential in energy management. This work is an effort to combine numerical (Energy Plus) and ML methods for energy services forecasting in a campus building. *RandomForest* predictor, in combination with feature selection techniques, is used owing to its ability to deal with complex data compared to other ML algorithms. Four ML models have been constructed taking the input from EP simulations and meteorological data for one year and predicting the energy service for next year in hourly fashion. Three type of errors (MAE, RMSE, and CV-RMSE) have been calculated and are used to compare the model performance against internationally accepted standards for hourly prediction. CV-RMSE being scale independent provides a good comparison between models and its value is less than 30% except HVAC model where it is 37%. Overall, the models performed significantly well and with further improvements can help in energy services forecasting with minimal error.

NOMENCLATURE

AI	Artificial Intelligence
ARMA	AutoRegressive-Moving Average
CV-RMSE	Coefficient of Variation of Root Mean Square Error
DT	Decision Tree
EP	Energy Plus
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine learning
MTLF	Medium-term load forecasting
RF	Random forest
RMSE	Root Mean Square Error
STLF	Short-term load forecasting
SVM	Support Vector Machine
VSTLF	Very short-term load forecasting

1. INTRODUCTION

With the increasing trend of the world population and the consequent increase of demand for energy services, the demand of energy in the world is increasing. Only in European Union, the buildings consume the 40% of total energy consumption [1]. On one hand, there is a massive research going on in the field of energy management that includes energy saving by shifting of least important tasks to off-peak hours and smart control of HVAC and lighting equipment. While on the other there a trend of passive housing and smart grids. The fundamental feature of smart grids is load forecasting which helps the operator to take effective and efficient decisions.

The buildings account for 30 % of global CO_2 emissions and 36 % of greenhouse gas (GHG) emissions only on European Union. These GHG emissions give rise to the atmospheric temperature and cause a devastating climate changing effect globally [2]. There is already an abundance of historical and meteorological data of buildings that needs to be utilized smartly to help shape the decarbonizing building strategies.

Load forecasting helps in several operating decisions such as management, planning, scheduling and load dispatching. An accurate result of load forecasting is highly desirable as the procedure takes lot of time and cost. It has been claimed in literature that just 1 % increase in the prediction error can cause a loss of millions of dollars every year [3]. Load forecasting has been categorized in four different categories depending upon their interval of forecasting. Short-term load forecasting (STLF) has major role in controlling the electricity price and demand close to real time, help schedule the fueling and other such operations.

On the other hand, Long-term load forecasting (LTLF) helps balancing the demand and production in case of smart grids or planning the energy policies. LTLF is much more complex compared to the STLF as it is affected by seasonal variation and uncertain future event changing the demand heavily. Medium-term load forecasting (MTLF) is useful maintenance scheduling, coordination of load dispatch and price settlement so that demand and generation is balanced.

There are numerous methods available for building demand forecast. All the available methods can be classified easily into three categories; Numerical, Analytical and predictive. Numerical methods include

TRNSYS, Energy Plus, DOE-2, etc. These modelling techniques need a considerable amount of real data and computation time to build the physical models for simulation of future consumption. Still it is hard to use them for online or real time applications, as it also requires the study of human energy utilization behavior [3]. Analytical model rely on in-depth knowledge of processes and the law governing them but they are advantageous to the numerical methods in terms once calibrated can be used anywhere.

Contrary to them predictive methods like Artificial Neural Networks (ANN), Decision Trees (DT), etc. are highly accurate and quick [4]. Random forests (RF) can be said the extension of DT´s as DT is their binary element. RF outsmarts the most of AI prediction techniques owing to its appealing characteristics which include [4]; (i) it´s interaction between predictors (ii) its basis of ensemble techniques allows it to learn the complex models (iii) it requires less hyper parameter tuning compared to its competitors.

The objective of this work is to combine the numerical and machine learning techniques to have better forecast results. Energy Plus software will be used for numerical part of work and RF machine learning technique for the Predictive part.

The EP simulations will be run on an already available physical model for a building of Tecnico´s alameda campus. The default weather data of EP software will be replaced with the real weather data available for campus´s open source meteorological platform. Simulation will be run for two years 2017 and 2018 to get the energy services.

The RF model will be trained for the one year´s (2017) data and the energy services will be forecasted for the next whole year (2018) in hourly fashion. The EP simulated data for 2018 will be used to validate the model.

The prediction will also be compared with the real data of power consumption for the building but only total energy consumption and HVAC consumption is available for this purpose.

Internationally accepted criteria for errors in hourly forecast will be used to check the fitness of used model.

The use of real weather data in EP simulations is supposed to give the better results compared to default weather data. The RF model is thought to produce less error when feature selection techniques will be used for prediction.

2. LITERATURE REVIEW

Time series load forecasting can be categorized in four groups depending upon the time intervals of forecasting; [5]

- Very short-term load forecasting (VSTLF) has the time period ranging from a few minute to half hour or few hour. Its aim is to adjust and control the demand and price in real time.
- Short-term load forecasting (STLF) has time period ranging from one day to one weak ahead and aims at economic dispatch and optimal generator unit commitment.
- Medium term load forecasting (MTLF) includes the forecasting from one month to one year ahead. Its purpose is to maintain the balance between generation and consumption for maintenance scheduling.
- Long-term load forecasting (LTLF) has the forecasting horizon longer than one year. It is necessary for the future electricity network planning conditions.

Short and medium term forecasting is important for economical operation, schedule the fuelling and maintenance operation while long term forecasting is useful for planning operation and capacity expansion [5]. STLF is slightly affected by weather conditions and the social behaviour of the community and in some cases, such algorithms report less than 1% mean absolute percentage error (MAPE) [6]. LTLF involves the uncertainties introduced by seasonal issues and distant future, which makes LTLF a challenging task [7].

Conventional methods of forecasting include statistical methods, which exhibit a white box model where the internal structure of process is well known and can be interpreted by using mathematical formulas and equations. This mathematical explanation allows understanding of the relation between input and outputs. Widely used statistical methods include multiple linear regression [8], autoregressive-moving average (ARMA) [9], auto-regressive integrated moving average (ARIMA) [10], Kalman filter [11], general exponential technique [12] and stochastic time series [13].

These white box statistical models are easily to implement and interpret but their shortcoming to handle non- linear and large data sets makes the machine learning methods need of the time.

There are several engineering methods and tools to forecast energy consumption in buildings but most of them are complex and require the physical models of the buildings or the space as well as the human behaviour and pattern of energy utilisation. DOE-2, Energy Plus, BLAST ESPr. are some of the tools used for energy efficiency and simulations for proper energy management.

In 1993-1994 American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) organized the first edition energy prediction contest, the Great Energy Predictor Shootout (GEPS) with the purpose of predicting energy consumption of commercial buildings in hourly fashion. The successful participant developed a machine-learning algorithm using a model based on sensors, which relied on domain knowledge of used building [14].

Following the GEPS, the black box type machine learning models became the topic of research because of their ability to learn and implement the complex patterns with minimal human interaction. Most of the times, the internal mechanism of these black box methods is unknown and difficult to interpret. The significant works include the use of Artificial Neural Network (ANN) by Chae et al. [15] for a next day prediction of electricity consumption in a commercial building with a time period of 15 minutes and Fu et al. [16] using Support Vector Machine (SVM) to forecast the coming day’s load forecast of a public building in Shanghai.

In a simplified definition, machine learning is a process of gathering all the available data, extracting the relevant information from it and developing a model which best explains the past and future datasets [17].

Decision tree (DT) is a widely used machine learning technique, which includes the classification and regression trees (CART) [18]. The decision Tree is a kind of an inverted tree where the top most node is called the root node, which has all the training data in it. Each decision node applies a test to the input data and the outputs are more than one, giving the different result in each case. The final nodes where a branch ends is called the leaf of tree. There are several leaves in a DT, which store the results of each case.

Random forest technique was first developed by Breiman in 2001 for both classification and regression [19]. The main purpose of RF was to overcome the shortcomings of DT by using multiple DT´s to generate a forest called RF. Since then it has been used in various fields for forecasting

Feature selection is a process of selecting a subset features and using them in model construction. Feature selection is the key in machine learning algorithms, which immensely affects the performance of the model. Some times in the data, there are numerous features and using all of them can lead to a dead end [20]. The features helps us to;

- Reduce overfitting to avoid the use of noisy data for prediction
- Improve accuracy by removing the misleading data
- Minimize the training time

Feature selection (FS) is a concept of vital importance in Machine Learning. Features are the inputs to any ML algorithm to train the model for learning the data and giving the best results of prediction. In real life algorithms, these features can range into hundreds and thousands. Therefore, a quest to find some key features which best represent the data and use them for the prediction, is must. Using all available feature leads to high computational cost, time and error [21].

Main categories of FS methods are filter, wrapper and embedded methods. Jurado et al. [22] applied the different building STLF techniques like Fuzzy Inductive Reasoning, RF and NNs using some filters for feature selection. These hybrid ML models were compared to statistical model ARIMA. The input data was for the houses in Catalonia, Spain. The results showed that Artificial Intelligence (AI) methods using FS step outperformed the statistical ones by giving 20 percent more accurate results.

Random forest is machine-learning technique belonging to ensemble methods. Ensemble are the methods of combining results of several different models working on same problem to get more flexible (less bias) and less sensitive (less variance) outcomes. The widely used ensemble methods are boosting and bagging [23].

Bagging works by training models in a parallel fashion where each tree or model uses a separate set of features to predict or classify the target variable. The outcome is the aggregate of all models used.

Boosting is rather a sequential process where the consequence of one model works as input to the next one, which helps learning from the mistakes made at earlier stage.

Random forest is a bagging type of ensemble learning methods. The binary unit of the random forest is decision tree. Decision tree is also an easy to implement approach of prediction but it results in high variance when the number of predictors (features) are more. Then random forest appears to help with the problem of high variance. One of the advantages of random forest is its ability to handle with large number of features [24].

3. METHODOLOGY

3.1. Data acquisition

The data to be used for analysis in this work is taken for a central building in Instituto Superior Tecnico (IST), Lisbon. This building is in the heart if IST as shown in **Error! Reference source not found.**. This central building has offices, classrooms, lecture halls, conference rooms and a library. The data is collected from the website[1] that is a public platform and it provide the weather data for mainland Portugal. This weather station is located on the rooftop of South Tower of campus at an altitude of 135 meters, with coordinates 38.736°N, 9.138°S. This station provides the following main weather elements with every five minutes.

- Temperature
- Humidity
- Atmospheric pressure
- Wind speed and direction
- Precipitation
- Total solar radiation on horizontal plane

The weather data for 2017 and 2018 is used in this work. There were some gaps in the data collected from this platform. The reason can be the fault in system or the absence of operator in public holidays. Almost of 90 days of data was missing across the time span of two years at different times of the year. This missing was then obtained from another public platform called CAMS[2] (Copernicus Atmosphere Monitoring Service), which provides the information related to air pollution and health, solar energy, greenhouse gases and climate forcing everywhere in the world.

Energy Plus is an open source building energy simulation software for modelling building heating, cooling, lighting, ventilating, and other energy flows.

In order to simulate the energy consumption, Energy Plus requires the two main files. One being the input file, which contains the building's physical description, installed equipment, lightings and the usage pattern of all those energy-consuming devices based on the real time scenarios of usage of those installed systems.

The second file needed for simulation is the weather file. Usually this file already in the software. It has the average of all weather entities in that file for last 30 years.

[1] http://meteo.tecnico.ulisboa.pt

[2] https://atmosphere.copernicus.eu

However, for this work, we needed the two energy consumption simulations with respect to real weather data for the specific year. Therefore, the default weather file is changed with the real weather file for both years 2017 and 2018.

Unfortunately, not all the required data was available for the building location and the only the global solar radiation was known. The horizontal and beam components of hourly solar radiation were computed using the concept of clearness discussed in book [25].

3.2. Pre-processing data

In-built python libraries are imported to the working notebook to pre-process the data. The main libraries used here are Pandas and NumPy, Matplotlib and Scikit-Learn. Pandas here helped us to read the data saved on computer in CSV format. NumPy performs scientific computing operations in the data analysis. Matplotlib helps visualizing data and comparing two data sates or entities in graphical from. It contains simple pyplot, bar-chart, histogram, correlation matrix, scatter plot to name a few. Scikit-Learn is the main machine-learning library in python. Pre-processing of data involves finding any anomalies or missing values in data. We have two CSV files two, one having real weather data for year 2017 and other having output of energy plus simulations for 2017. Following steps summarize the pre-processing of data for this work;

- Converting temperature units from Celsius to kelvin to avoid any errors created by negative temperatures
- Dropping all the columns of Energy Services 2017 file except the one target variable (Facility Energy consumption)
- Converting Facility energy consumption from Joules to Kilowatts to easily understand the data
- Bringing both data sets in python time series format
- Merging data sets on time columns using pandas library

3.3. Feature creation and selection

Two sets of features have been used for the model training. First set has nine features is provided by real weather data.

- Temperature
- Relative Humidity
- Atmospheric pressure
- Wind speed
- Wind gust
- Precipitation (hourly and daily)
- Solar radiations on horizontal plane (beam, diffused and global)

The second set contains constructed features using engineering knowledge and exploratory data analysis in python. It contains six features;

- Hour of day
- Day of week
- Month of the year
- Consumption in previous hour
- Average consumption of previous three hours
- Type of the day

As none of the single method is efficient for feature selection, so both filter and wrapper methods have been used for feature selection, and the top features in all methods have been used for the model training.

Three of the filter methods used are namely feature scores, feature importance and correlation matrix. Recursive Feature elimination is used here from wrapper methods for feature selection.

Finally, eight features, which are common in all above methods, are selected for the model training. The 10 best features are as follows;

1. Direct normal radiation
2. Diffused Horizontal radiation
3. Global Horizontal radiation
4. Temperature
5. Holiday
6. Hour
7. Relative humidity
8. Previous hour consumption

3.4. Model training, tuning and testing

Setting up the model starts by splitting the data into training and testing sets. There is inbuilt function of *train_test_split* in Scikit learn library of python which splits data automatically in 75% training -25% testing.

RandomForestRegressor from ensemble methods in Scikit library is imported for training and it has to be tuned for best results.
Hyperparameter tuning means running and fitting the model with many different values of parameters to find the optimum value of each of them, which provides the best results of model.

Here a technique cross validation CV helps optimizing these parameters. Cross validation further divides the training set into K number of subsets (we provide the K value). The model uses K-1 subsets for training the model and the K^{th} set for testing which is called cross validation technique.

Trying all the combinations consumes a significant amount of computation time. To avoid this *RandomizedSearchCV* command is used where the model does not try every combination of parameters but only a specific number of parameters set by the user. This information is passed under *n_iter* parameter of *RandomizedSearchCV.* A value of 100 is used for this work along with 10-fold CV that means the models is run and fit with 1000 random combinations and returns the best combination. The optimized set of parameters is shown in Figure 1.

```
{'n_estimators': 1400,
 'min_samples_split': 2,
 'min_samples_leaf': 4,
 'max_features': 'sqrt',
 'max_depth': 80,
 'bootstrap': False}
```

Figure 1 optimized parameters for model training

3.5. Error calculation

Prediction performance is evaluated by using three metrics that are mean absolute error (MAE), root mean square error (RMSE), and coefficient of variance of root mean square error (CV-RMSE). First three are scale dependent while the last one is independent [26]. The three metrics are mathematically expressed by equations;

$$MAE = \frac{\sum_{k=1}^{n} |y_a - y_p|}{n} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n} (y_a - y_p)^2}{n}} \quad (2)$$

$$CV-RMSE = \frac{\sqrt{\frac{\sum_{k=1}^{n} (y_a - y_p)^2}{n}}}{\frac{\sum_{k=1}^{n} y_a}{n}} \quad (3)$$

Where y_a means actual value and y_p means predicted values and n is total number of values in data set.

4. RESULTS AND DISCUSSION

There are two types of results. First, the results about the simulation from EP software are discussed followed by the ML model performance results.

4.1. Simulation results

Energy Plus software uses the default weather file and it had to be replaced with the real weather data file for both the years 2017 and 2018. The simulation results have shown that it improves the output by providing values closer to the actual power consumption.

It has been found out that using default weather file for EP oversimplifies the model and gives the results that are far away from the actual power consumption. Therefore, simulations have been run twice, using default and real weather files and the simulation results compared to the actual values. The graphical analysis has shown that the real data provides the values closer to the actual ones specifically for the months of March, April, November and December. Figure 2 and Figure **3** show the concerned part of graphs zoomed in.

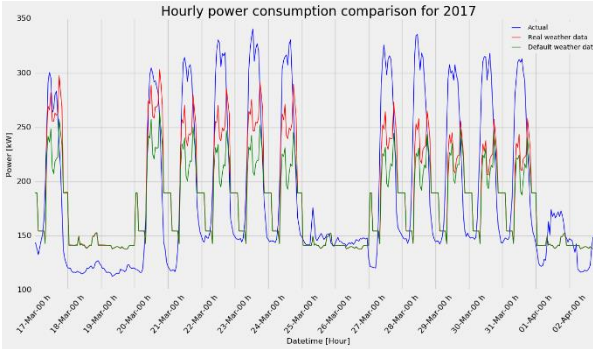

Figure 2 EP Simulations comparison for 2017 data

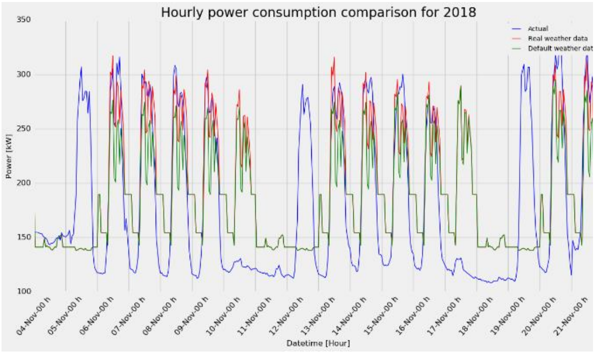

Figure 3 EP simulations comparison for 2018 data

4.2. Forecasting results

Power consumption is forecasted for four types of consumption.

- Facility
- Building
- HVAC
- Exterior lights

They will be discussed one by one. The prediction results for facililty power consumption compared to EP simulations and actual consumption are shown in Figure 4.

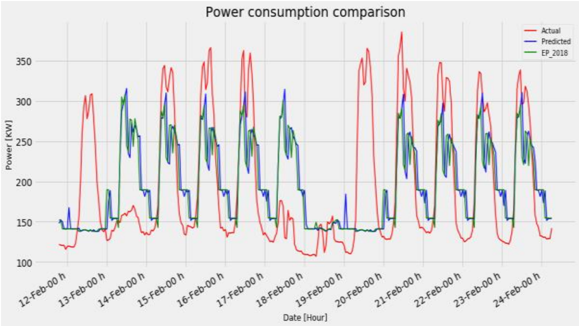

Figure 4 RF model compared against EP simulation and actual power consumption for 2018

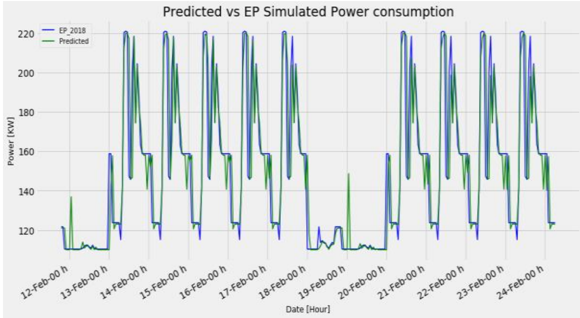

Figure 6 RF model prediction compared against EP simulations for 2018

The validation metrics used here are MAE, RMSE and CV-RMSE which gives the values 8.6 KW, 16.7 KW and 9.2% respectively.

Comparing the prediction with actual consumption gives the values of metrics (MAE, RMSE and CV-RMSE) 52 KW, 69 KW and 38% respectively. These high numbers indicate a faulty model for prediction but this hypothesis is proven wrong by running the model with actual power consumption as input and comparing it with actual power consumption for 2018. The values of MAE, RMSE and CV-RMSE comes out to be 7.2 KW, 11.4 KW and 6.3% respectively which are quite reasonable and well in range. This is graphically represented in Figure 5.

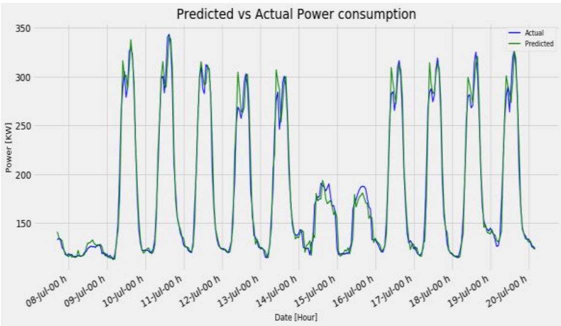

Figure 5 RF model performance using real data of power consumption

Building power consumption comprises of interior equipment and interior lights. The prediction is graphically compared to the EP simulated values shown in Figure 6 that shows a good fit. The metric MAE, RMSE and CV-RMSE has the values 6.5 KW, 13.4 KW and 9.5% respectively.

HVAC consumption in EP consists of the heating and fans power consumption. The forecasting results compared to the EP simulations and actual HVAC consumption are shown in Figure 6. The metrics MAE, RMSE and CV-RMSE comes out to be 4 KW, 8.8 KW and 37% respectively.

The next step was to train the model with the real HVAC consumption in 2017 and test it against the real HVAC consumption in 2018. The metric MAE, RMSE and CV-RMSE has the values 3.5 KW, 6.4 KW and 32% respectively which are better than the model tested on simulated data. The trend is compared in Figure 6 that shows the better fit compared to the simulated data and proves the accuracy of the model.

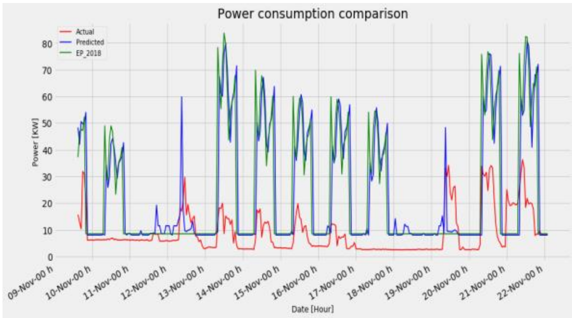

Figure 7 RF model performance compared against actual HVAC consumption for 2018

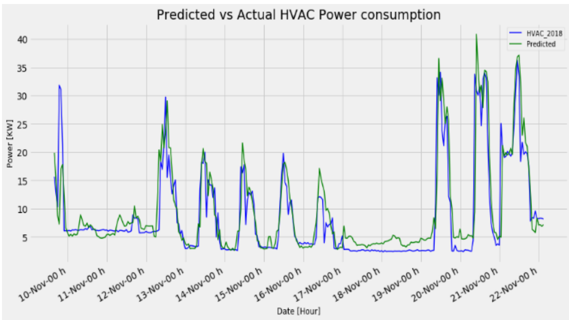

Figure 8 RF model performance with real HVAC consumption data

For exterior lights, the comparison with the EP simulation results is shown in Figure 9 that shows the good fit. The metrics MAE, RSME and CV-RMSE has values of 0.29 KW, 0.81 KW and 28% respectively.

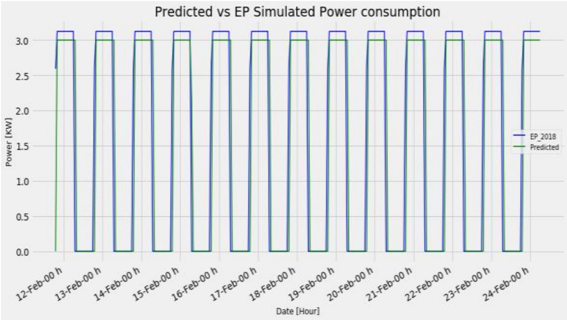

Figure 9 RF model performance compared against EP simulations for 2018

Overall, the model performed well. There were some errors exceeding the limit but they are justified by the vacation month (August) false prediction especially for HVAC prediction. Table 1 summarizes the errors in accordance with the validation limits. Here RF model errors are comparing prediction with simulation results.

Table 1 All four RF model errors compared to international standards

	CV-RMSE [%] for hourly prediction			
	RF model	IPMVP	FEMP	ASHRAE
Facility	9.2	20	30	30
Building	9.8			
HVAC	32			
Ex .Lights	28			

5. Summary

This work combined the numerical and predictive ML methods of forecasting energy services in the buildings. It was found out that putting an effort to collect and prepare the real weather data for EP simulations helped go get better prediction results. The predicted consumption with real weather was closer to the real values for first and third quarter of the year.

RandomForestRegressor ML model was applied to forecast the energy services based on EP simulated data. The other input to the model were metrological data for the first year and some features based on historical and engineering knowledge. The results showed that the historical data like consumption in previous hour has a close relation to the output. Feature Selection step resulted in more accurate results and saved the computational time.

Hyperparameter tuning step with using *RandomizedSearchCV* and 10-fold cross validation was implemented to obtain optimized parameters for *RandomForestRegressor* model. Although it consumed a significant amount of time but it helped in shaping the model performance.

The ML model was validated using three types of errors; MAE, RMSE (scale dependent) and CV-RMSE (scale independent). The percentage error (CV-RMSE) was compared against the limits defined by the internationally accepted organizations for the hourly building energy prediction.

The model performed significantly well with the prediction error of 15 percent of less in case of Exterior lights, Facility and Building energy consumption using the EP simulated data. Using the available actual data for total energy consumption resulted in approximately 10 percent error (even less).

The model performed somewhat strange in case of HVAC consumption prediction. The percentage error was more than the limits defined internationally (37 % for EP simulated and 32 % for actual data).

6. Future work

The suggestions for the future work are to use state of the art models for the simulations and use the latest version of EP software to avoid the unnecessary conversions and save the time. The other ML models, which are out of the scope of this work, can be used to compare the model performance and get better forecasting results.

7. REFERENCES

[1] "Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energyperformance of buildings," 2010.

[2] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy Build.*, vol. 147, pp. 77–89, Jul. 2017.

[3] G. F. Fan, L. L. Peng, W. C. Hong, and F. Sun, "Electric load forecasting by the SVR model with differential empirical mode decomposition and auto regression," *Neurocomputing*, vol. 173, pp. 958–970, Jan. 2016.

[4] M. W. Ahmad, M. Mourshed, B. Yuce, and Y. Rezgui, "Computational intelligence techniques for HVAC systems: A review," *Build. Simul.*, vol. 9, no. 4, pp. 359–398, Aug. 2016.

[5] A. Yang, W. Li, and X. Yang, "Short-term electricity load forecasting based on feature selection and Least Squares Support Vector Machines," *Knowledge-Based Syst.*, vol. 163, pp. 159–173, Jan. 2019.

[6] H. M. Al-Hamadi and S. A. Soliman, "Long-term/mid-term electric load forecasting based on short-term correlation and annual growth," *Electr. Power Syst. Res.*, vol. 74, no. 3, pp. 353–361, Jun. 2005.

[7] T. Al-Saba and I. El-Amin, "Artificial neural networks as applied to long-term demand forecasting," *Artif. Intell. Eng.*, vol. 13, no. 2, pp. 189–197, Apr. 1999.

[8] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical

consumption: A machine learning case study," *Energy Build.*, vol. 49, pp. 591–603, 2012.

[9] S. S. Pappas, L. Ekonomou, D. C. Karamousantas, G. E. Chatzarakis, S. K. Katsikas, and P. Liatsis, "Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models," *Energy*, vol. 33, no. 9, pp. 1353–1360, Sep. 2008.

[10] A. Veit, C. Goebel, R. Tidke, C. Doblander, and H. A. Jacobsen, "Household electricity demand forecasting - Benchmarking state-of-the-art methods," *e-Energy 2014 - Proc. 5th ACM Int. Conf. Futur. Energy Syst.*, pp. 233–234, 2014.

[11] H. Takeda, Y. Tamura, and S. Sato, "Using the ensemble Kalman filter for electricity load forecasting and analysis," *Energy*, vol. 104, pp. 184–198, Jun. 2016.

[12] W. Christiaanse, "Short-Term Load Forecasting Using General Exponential Smoothing," *IEEE Trans. Power Appar. Syst.*, vol. PAS-90, no. 2, pp. 900–911, Mar. 1971.

[13] Y. Chakhchoukh, P. Panciatici, and L. Mili, "Electric Load Forecasting Based on Statistical Robust Methods," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 982–991, Aug. 2011.

[14] D. J. C. MacKay, "Bayesian Non-Linear Modeling for the Prediction Competition," in *Maximum Entropy and Bayesian Methods*, Dordrecht: Springer Netherlands, 1996, pp. 221–234.

[15] Y. T. Chae, R. Horesh, Y. Hwang, and Y. M. Lee, "Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings," *Energy Build.*, vol. 111, pp. 184–194, Jan. 2016.

[16] Y. Fu, Z. Li, H. Zhang, and P. Xu, "Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices," in *Procedia Engineering*, 2015, vol. 121, pp. 1016–1022.

[17] O. Pauly, "Random Forests for Medical Applications," *Basic Appl. Ecol.*, vol. 4, no. 5, pp. 441–451, 2003.

[18] A. Troncoso, S. Salcedo-Sanz, C. Casanova-Mateo, J. C. Riquelme, and L. Prieto, "Local models-based regression trees for very short-term wind speed prediction," *Renew. Energy*, vol. 81, pp. 589–598, Sep. 2015.

[19] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[20] Raheel Shaikh, "Feature Selection Techniques in Machine Learning with Python," *towardsdatascience.com*, 2018. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Accessed: 24-Oct-2019].

[21] S. Salcedo-Sanz, L. Cornejo-Bueno, L. Prieto, D. Paredes, and R. García-Herrera, "Feature selection in machine learning prediction systems for renewable energy applications," *Renew. Sustain. Energy Rev.*, vol. 90, pp. 728–741, Jul. 2018.

[22] S. Jurado, À. Nebot, F. Mugica, and N. Avellana, "Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques," *Energy*, vol. 86, pp. 276–291, Jun. 2015.

[23] Lujing Chen, "Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained," *towardsdatascience.com*, 2019. [Online]. Available: https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725. [Accessed: 24-Oct-2019].

[24] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, Jun. 2012.

[25] J. A. Duffie and W. A. Beckman, *Solar engineering of thermal processes*, 4th editio. Wiley, 2013.

[26] C. Fan, F. Xiao, and Y. Zhao, "A short-term building cooling load prediction method using deep learning algorithms," *Appl. Energy*, vol. 195, pp. 222–233, Jun. 2017.