



A statistical and predictive modeling study to analyze impact of seasons and covid-19 factors on household electricity consumption

Gaikwad Sachin Ramnath 

Department of Electronics and Telecommunication, Symbiosis Institute of Technology (SIT), Symbiosis International Deemed University (SIDU), Pune, India, sachin.r.gaikwad@outlook.com

Harikrishnan R 

Department of Electronics and Telecommunication, Symbiosis Institute of Technology (SIT), Symbiosis International Deemed University (SIDU), Pune, India, rhareish@gmail.com

Submitted: 06.05.2021

Accepted: 01.11.2021

Published: 31.12.2021



Abstract: Load is dynamic in nature and changing from aggregated load to disaggregated loads. Hence, need to analyze individual household's energy consumption pattern. Many factors are contributing to household electricity consumption (HEC). The most influencing factor is the end user's behavioral aspect. The calendar and seasonal factors are directly affecting user's behavior activities. This paper consists of two aim, first aim is to validate the performance of traditional predictive models and second aim is to identify the best-fitted predictive model from five predictive models namely: Random Forest, Linear Regression, Support Vector Machine, Neural Network (NN) and Adaptive Boosting. The orange tool is used to simulate the predictive models. The JASP tool is used for statistical analysis of the dataset. From the predictive modeling study, the NN model is the most fitted model. The values of the performance matrix parameter like MSE, RMSE and MAE of the NN model is observed to be 0.558, 0.747 and 0.562 respectively. This study gives insights to researchers and utility companies about traditional predictive models that can predict the HEC under anomaly situations like Covid-19. This study also helps the researchers in using Orange and JASP tool to perform the statistical and predictive modeling.

Keywords: *Calendar effect, Household electricity consumption, Predictive modeling, Seasonal factor, Statistical analysis.*

Cite this paper as: Ramnath, G, S., Harikrishnan, R., A statistical and predictive modeling study to analyze impact of seasons and covid-19 factors on household electricity consumption. *Journal of Energy Systems* 2021; 5(4): 252-267, DOI: 10.30521/jes.933674

1. INTRODUCTION

The electricity supply is a fundamental service that is required for every nation for overall development. In the United States, the electricity consumption of the residential and commercial building was around 73% and 41%, respectively in the year 2015, whereas, China, residential energy use has been extended around 13.6% of the total electricity use in the year 2017 [1]. In the US and European countries along with the electricity consumption, the CO₂ emissions is also increased, which is around 40% of electricity consumption, leads to 36% of CO₂ emissions. The world-wide dwelling energy demand is continuously increasing and it is around 20% to 40% with 33% of Green House Gas (GHG) emission. Moreover, the energy consumption share of the household is nearly 20%, among the total use of energy [2]. So, the major usage of electricity supply is in residential, commercial, industrial and agricultural sectors. Among these, less research is on the residential sector on how the end-user is using power. The main reasons are the diverse factors on which power consumption depends, and lack of availability of quality, sufficient and diverse household dataset. Moreover, household energy uses observed variations in statistical and sociological analysis. The inconsistency of energy utilization is a challenge to statistical modeling, which is due to limited data set on the variability of usage. Furthermore, the sources of this deviation of present household energy use are largely unknown. Electricity consumption depends on many factors, a few of them are, geographical location, physical characteristics of building and household, type of appliance ownerships, inside and outside weather conditions, socio-demographics, calendar and seasonal effects, behavior of the end user and occupant, psychology, sociology, and culture [3,4,5,6].

For energy optimization and demand-side management, the exact load demand from the distribution side on a different time scale has to be known. One of the influencing parameters on energy use is the behavior of the consumer. The user behavior is closely related to the calendar and seasonal variability. So, the calendar events and different periods are one of the significant aspects to understand the consumption pattern of the end user. This paper aims to apply the statistical analysis technique with different methods to understand the impact of the calendar and seasonal factors on household energy consumption patterns. The detailed analysis is done based on calendar periods such as Covid-lockdown, vacation, weekday, and weekend [7]. Furthermore, calendar information with weekly and seasonal datasets is also important in load prediction modeling. A literature study shows that the performance of the prediction model is better when the calendar effect is included. Moreover, due to calendar effects, it is possible to capture the behavior of the end user. The calendar information is also essential to determine the shape of the load profile of the household [8,9]. The author [10] performed an experiment, based on seasonal variations in the demands of the household loads. The study used two data sets, one being the monitored dataset having data gathered from 58 English households between July 2011 to December 2011 and the other being the synthetic data set generated by using a time-of-use based load modeling tool.

This study revealed that there is a significant impact of seasonal and calendar data on the HEC. This paper consists of two aims, first aim is to contribute how the seasonal and calendar factors are influencing the HEC through literature study and obtained results. The second aim is to shed some light on how the existing statistical method and predictive models are performing in Covid -19 pandemic situation for energy prediction. To achieve this, the paper used Orange tool-based five traditional predictive learning algorithms namely: RF, LR, SVM, NN and AB. Then compared and selected the best-fitted model based on the performance using various well-known indicators and metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Square Error (MSE). In addition, this study gives insights to the researchers and utility companies about traditional predictive models that can predict the HEC under anomalous situations like Covid-19. This study also helps the researchers in using Orange and Jeffrey's Amazing Statistics Program (JASP) tool to perform statistical and predictive

modeling. This paper quotes a future scope where this work can get extended by including holidays and outside weather conditions.

The paper consists of five sections. The first section is the introduction with the motivation of the topic. The second section discusses the data from its data collection to its interpretations. The third section represents the JASP tool-based statistical method and results in analysis. Similarly, the fourth section includes the Orange tool-based five learning algorithms based predictive modeling and result in analysis for the understanding of Calendar, Seasonal and Covid-19 effects on household electricity consumption. The fifth section discusses the conclusion and future directions.

2. THE DATA

The source of the dataset is from the Kaggle database (<http://www.kaggle.com/srinuti/residential-power-usage-3years-data-timeseries>) which is for around four years and one month (1st June 2016 to 7th July 2020). The data set is of time series and labelled type. This database is required for statistical analysis to understand the pattern of energy consumption in different calendar periods. The power usage dataset is from one house with two floors, which are located in Houston, Texas, USA. The same location has round ten months of summer (February to November) and two months of winter (December to January) seasons. The dataset is also providing information related to appliance ownerships and their usage in day and night time. The appliance usage in the daytime is security Digital Video Recorder (DVR) and POI cameras, two refrigerators, two 189.271 liters' water heater, and in the night time several electrical bulbs, television, washing machine, dryer and air conditioner running from evening 6.00 PM to morning 8.00 AM are used. For a better understanding of the dataset and to understand the results properly, it is important to know the behaviors or nature of usage patterns. In the Covid-lockdown period the Air Conditioner (AC), laptop, monitors, etc. are used during the daytime. During the vacation period, AC and electric bulbs are not in use for a whole day.

The use of AC is more at weekends in a year. Moreover, in the summer season, for cooling the room, the temperature is set to 25.56 °C, and during the winter season for heating the room, the temperature is set to 20 °C. The weekday is considered from 7.00 AM to 5.00 PM as working hours with set temperature in summer is 28.89 °C and during winter for heating set temperature is 15.56 °C. The dataset includes four variables namely, start date, value (kWh), day of week, and notes. The day of the week is ranging from 0 to 6. The last variable is noted which includes four labelled calendar events namely Covid-lockdown, vacation, weekends, and weekdays. The statistical analysis aims to understand the effects of calendar datasets with the seasonal factors on household power usage. This analysis applied manual classification technique using Microsoft Excel tool to classify different notes category. This classification needs to understand the effects of the calendar aspect on household electricity consumption.

3. STATISTICAL METHOD AND RESULT ANALYSIS

The JASP tool (<https://jasp-stats.org/>), is a free, open-source statistical package. This tool includes descriptive statistics which is mostly applied for data interpretation and data visualization [11].

Descriptive statistical analysis is one of the statistical techniques of the JASP tool. On the input side, a categorical variable can be split and analyzed using different plots and statistical approaches. The main requirement for the descriptive statistical technique is that the applied data should have the maximum continuous variables. The descriptive statistical technique consists of three main types of analysis namely: frequency tables, plots and statistics as shown in Fig. 1. The frequency table displays the

frequency of each variable. Moreover, three main types of plots are distribution (display density), correlation (histogram and scatter plot) and box (label outliers and color).

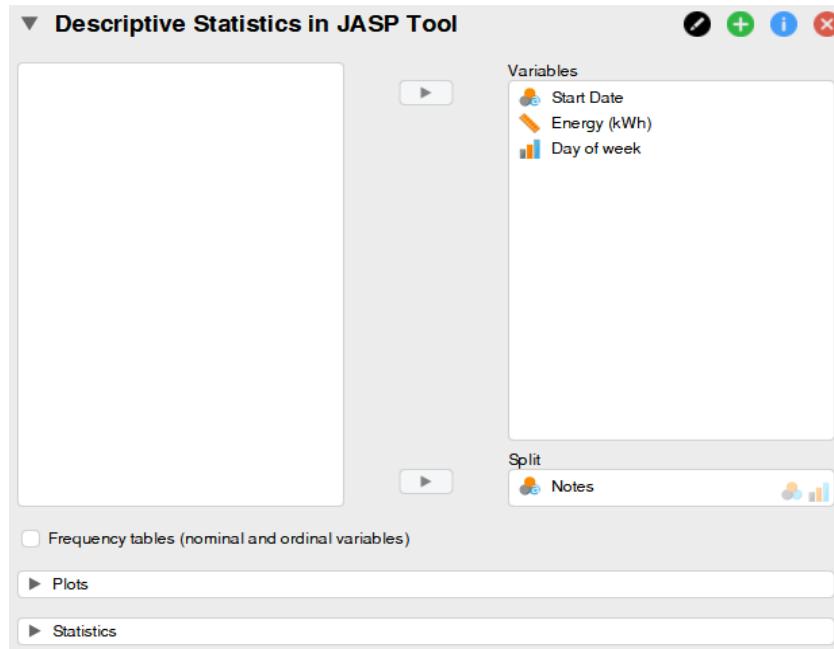


Figure 1. Input window of JASP tool (<https://jasp-stats.org/>)

Besides, the statistical part consists of four types of analysis namely: percentile values (quartiles and percentiles), central tendency (Mean, Median, Mode, Sum), dispersion (standard error of the mean, standard deviation, median absolute deviation, interquartile range, variance, range, minimum, maximum) and distribution (skewness, kurtosis and Shapiro-Wilk test) (<https://jasp-stats.org/wp-content/uploads/2020/11/Statistical-Analysis-in-JASP-A-Students-Guide-v14-Nov2020.pdf>) [11].

3.1. Statistical Result Analysis Including Whole Dataset

Table 1, shows the descriptive statistical analysis based on calendar notes such as lockdown (LDN), vacation (VCN), weekend (WND) and weekday (WKD) and energy consumptions in different notes.

Table 1. Descriptive statistical analysis

Parameters	Day of week				Value (kWh)			
	LDN	VCN	WED	WND	LDN	VCN	WED	WND
Valid	2305	1133	23299	9215	2305	1133	23299	9215
Mean	3.020	3.105	1.999	5.503	0.864	0.436	0.893	0.952
Std. Error of Mean	0.042	0.059	0.009	0.005	0.016	0.012	0.006	0.010
Median	3.000	3.000	2.000	6.000	0.547	0.275	0.501	0.531
Mode	5.000	4.000	3.000	6.000	0.323	0.182	0.292	0.283
Std. Deviation	2.026	1.992	1.410	0.500	0.747	0.390	0.915	0.956
Variance	4.104	3.967	1.989	0.250	0.558	0.152	0.837	0.913
Skewness	-0.020	-0.083	-3.038e-4	-0.011	1.780	2.662	2.156	1.966
SE of Skewness	0.051	0.073	0.016	0.026	0.051	0.073	0.016	0.026
Kurtosis	-1.294	-1.253	-1.294	-2.000	2.792	9.350	4.639	3.650
SE of Kurtosis	0.102	0.145	0.032	0.051	0.102	0.145	0.032	0.051
Range	6.000	6.000	4.000	1.000	4.035	3.077	6.382	5.323
Minimum	0.000	0.000	0.000	5.000	0.214	0.112	0.064	0.147
Maximum	6.000	6.000	4.000	6.000	4.249	3.189	6.446	5.470

3.1.1. Statistical analysis with calendar notes

The dataset consists of a total of 35952 hours. In the total dataset, 65% data in hours is on weekdays. Moreover, the data available for the vacation period is 3.15% and for Covid-lockdown is 6.4% which is very low compared to other calendar periods. The highest mean for energy consumption is observed on a weekend that is 0.952 kWh and the lowest mean is observed on vacation that is 0.436 kWh. The least median for energy consumption is observed in the vacation with 0.275 kWh and the highest median is observed in the lockdown period with 0.547 kWh. The standard deviation for energy consumption on a weekend is the highest at 0.956 kWh and it is the lowest on vacation with 0.390 kWh. The minimum energy consumption in different calendar periods is observed to be in the winter season with an exception of a weekday in the summer season. Similarly, the maximum energy consumption in different calendar periods is observed in the summer season. Moreover, the energy consumption range parameter is found to be greater on a weekday and lesser on vacation.

3.1.2. Density and counts of calendar notes

On an hourly basis mostly 0.5 kWh unit energy consumption is occurred in all calendar durations except the vacation period, as shown in Fig. 2 to Fig. 5 density Vs value graphs.

The highest hourly energy data consumed is from weekdays as compared to other calendar durations, as shown in Fig. 6 to Fig. 9 count graphs. It can be observed that there is a uniformity in all day counts energy consumption in weekdays and weekends days compared to Covid-lockdown and vacation.

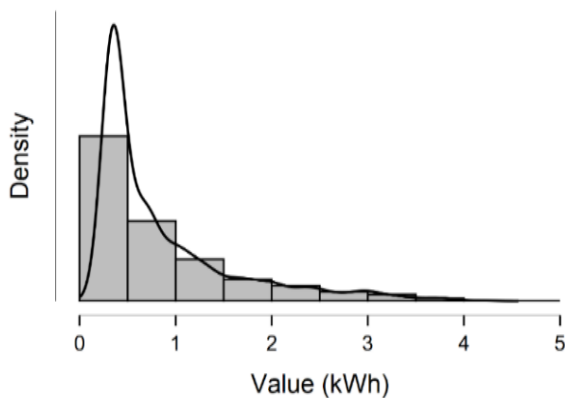


Figure 2. Units in covid-lockdown

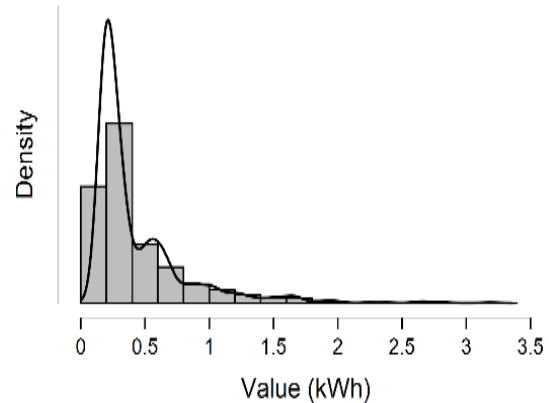


Figure 3. Units in vacation

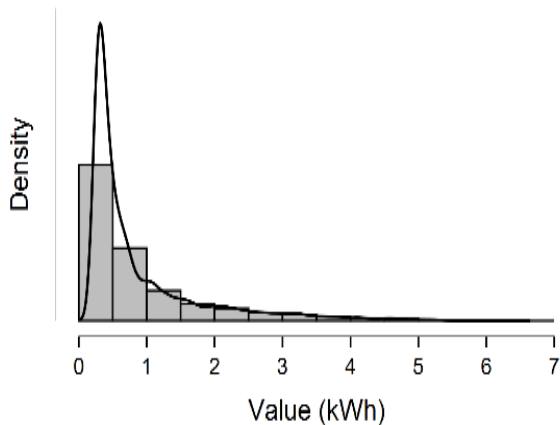


Figure 4. Units in weekdays

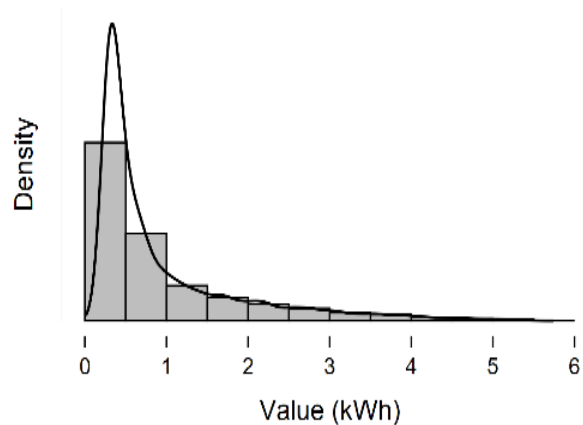


Figure 5. Units in weekends

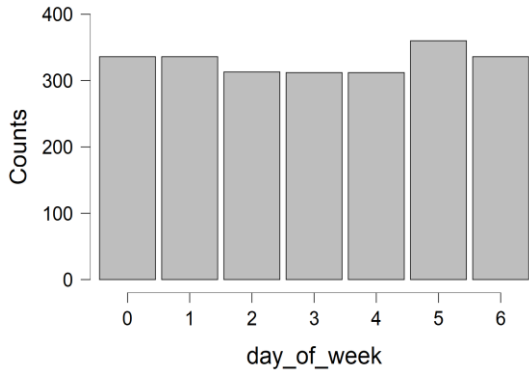


Figure 6. Day of week in covid-lockdown

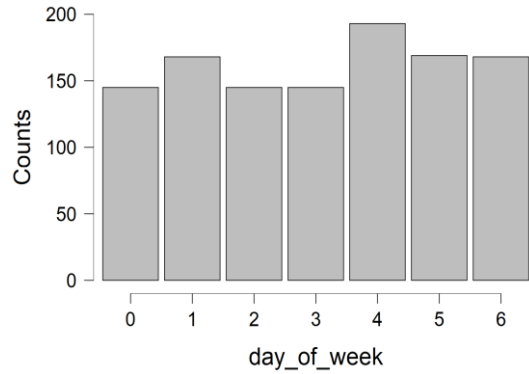


Figure 7. Day of week in vacation

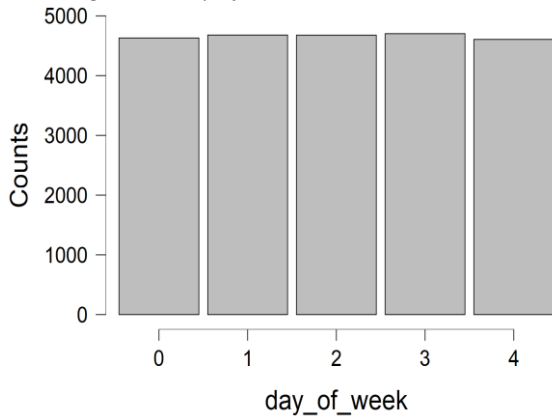


Figure 8. Day of week in weekdays

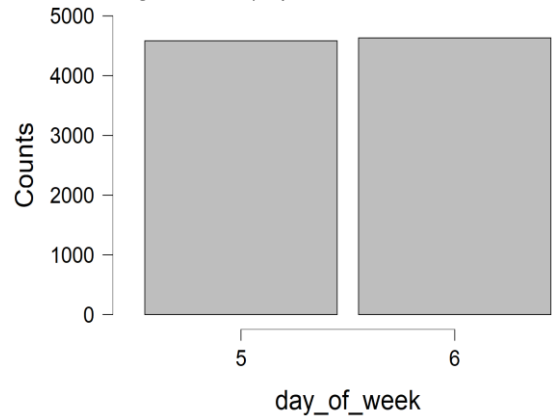


Figure 9. Day of week in weekends

3.1.3. Comparative study of notes

The boxplots comparative analysis of notes considering the day of week and value are shown in Fig. 10 and Fig. 11. The day of the week is uniformly distributed except for vacation duration. The maximum power consumption means of 0.952 kWh is observed on the weekends and the minimum mean of 0.436 kWh is observed in the vacation period.

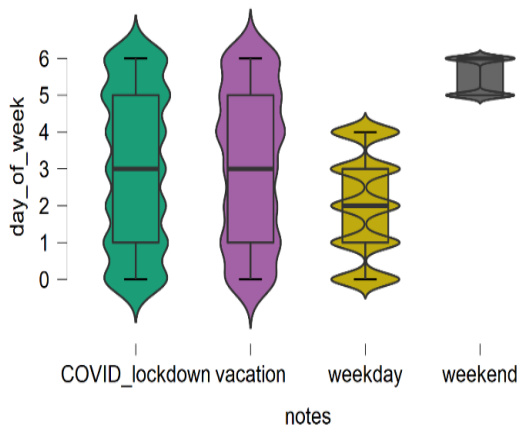


Figure 10. Comparison of notes with day of week

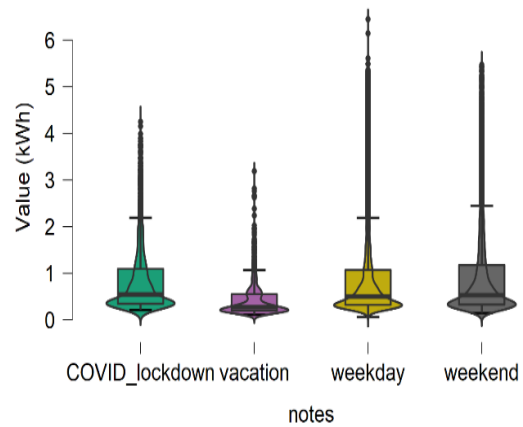


Figure 11. Comparison of notes with value (kWh)

3.1.4. Power consumption at notes

The Scatter plot analysis, that is drawn between a day of the week and value (kWh) are shown in Fig. 12 to Fig. 15. This plot shows how the input data set is scattered with its density. On weekdays and weekends, the power consumption pattern is observed to be the same and steady for all days. As on lockdown and vacation due to fixed schedule of work, the power consumption is more dynamic as shown

in Fig. 12 and Fig. 13. Moreover, around 1 kWh of minimum power consumption per day is observed in all notes except vacation.

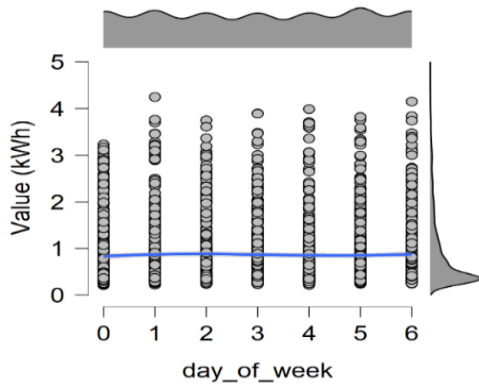


Figure 12. Units in lockdown

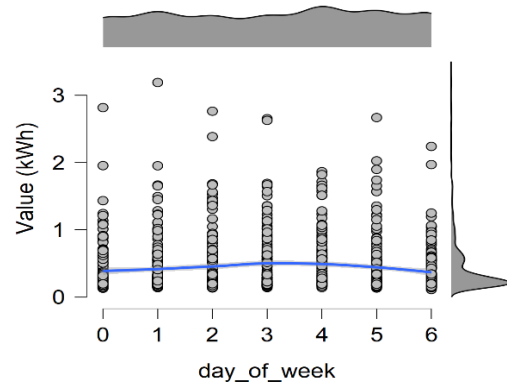


Figure 13. Day of week in vacation

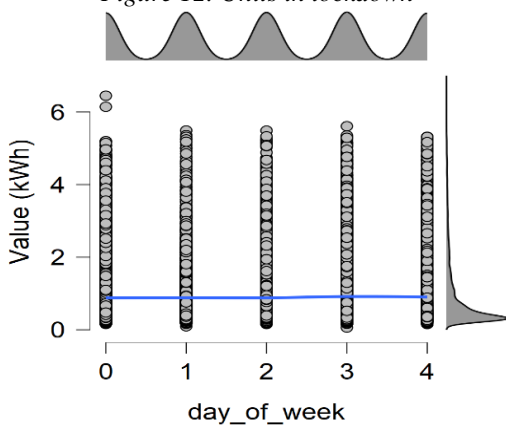


Figure 14. Units in weekday

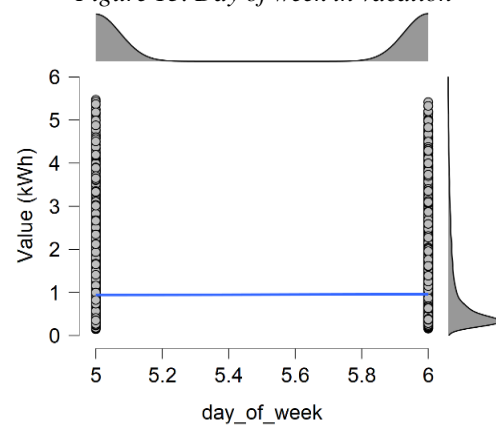


Figure 15. Day of week in weekend

3.2. Statistical Result Analysis Including Seasonal Effects

3.2.1. Effects of summer season on notes

The duration of the summer season is from February to November in a year and the data collection is done on households located in Houston, Texas, USA. The descriptive statistical analysis for the summer season is given in Table 2. The maximum power consumption shown in Table 1 and Table 2 is the same. It means that in the whole data set the peak power consumption is recorded in the summer season.

Table 2. Summer season descriptive statistical analysis

Parameters	Day of week				Value (kWh)			
	LDN	VCN	WED	WND	LDN	VCN	WED	WND
Valid	2232	580	20277	7967	2232	580	20277	7967
Mean	3.054	3.166	2.000	5.500	0.878	0.593	0.958	1.027
Std. Error of Mean	0.043	0.082	0.010	0.006	0.016	0.019	0.007	0.011
Median	3.000	3.000	2.000	6.000	0.567	0.475	0.533	0.585
Mode	5.000	3.000	1.000	6.000	0.323	0.210	0.292	0.310
Std. Deviation	2.029	1.973	1.412	0.500	0.754	0.468	0.960	1.002
Variance	4.117	3.893	1.993	0.250	0.569	0.219	0.921	1.004
Skewness	-0.034	-0.106	8.728e-5	-2.511e-4	1.740	1.972	1.964	1.771
SE of Skewness	0.052	0.101	0.017	0.027	0.052	0.101	0.017	0.027
Kurtosis	-1.302	-1.160	-1.297	-2.001	2.622	5.030	3.671	2.772
SE of Kurtosis	0.104	0.203	0.034	0.055	0.104	0.203	0.034	0.055
Range	6.000	6.000	4.000	1.000	4.024	3.022	6.382	5.308
Minimum	0.000	0.000	0.000	5.000	0.225	0.167	0.064	0.162
Maximum	6.000	6.000	4.000	6.000	4.249	3.189	6.446	5.470

3.2.2. Comparative study of notes in summer season

The box plot comparative analysis of notes with value and day of the week are shown in the Fig. 16 and Fig. 17. Weekday recorded a peak power consumption of around 6.446 kWh. A maximum standard deviation of 2.029 is observed in the Covid-lockdown period for a week. Fig. 17, shows the maximum variance of 4.117 occurred in Covid-lockdown and minimum variance of 0.250 occurred at the weekend.

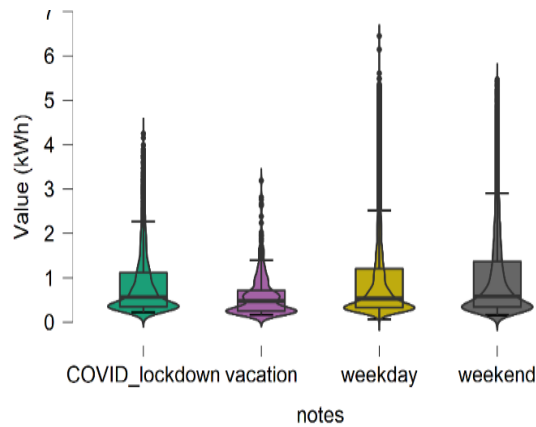


Figure 16. Comparison of notes with value (kWh)

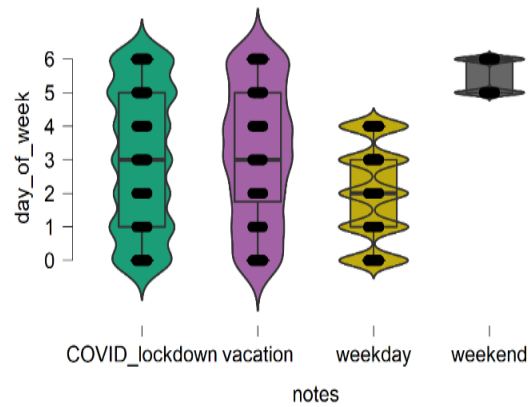


Figure 17. Comparison of notes with day of week

3.2.3. Winter season analysis

Table 3, shows the winter season analysis. The sample size of the winter season is less compared to the summer season. The reason behind this is that the winter season is for only two months in a year. The maximum standard deviation of energy consumption is observed on a weekend which is 0.252 kWh.

Table 3. Winter season analysis

Parameters	Day of week				Value (kWh)			
	LDN	VCN	WED	WND	LDN	VCN	WED	WND
Valid	73	553	3022	1248	73	553	3022	1248
Mean	2.000	3.042	1.992	5.519	0.433	0.272	0.455	0.469
Std. Error of Mean	0.191	0.086	0.025	0.014	0.025	0.007	0.004	0.007
Median	2.000	3.000	2.000	6.000	0.378	0.206	0.388	0.398
Mode	2.000	1.000	3.000	6.000	0.252	0.184	0.290	0.340
Std. Deviation	1.633	2.011	1.401	0.500	0.210	0.169	0.229	0.252
Variance	2.667	4.044	1.962	0.250	0.044	0.029	0.053	0.064
Skewness	0.000	-0.057	-0.003	-0.077	2.018	3.620	1.761	2.562
SE of Skewness	0.281	0.104	0.045	0.069	0.281	0.104	0.045	0.069
Kurtosis	-1.499	-1.345	-1.277	-1.997	6.072	20.654	5.048	13.883
SE of Kurtosis	0.555	0.207	0.089	0.138	0.555	0.207	0.089	0.138
Shapiro-Wilk	0.795	0.908	0.891	0.636	0.817	0.654	0.852	0.806
Range	4.000	6.000	4.000	1.000	1.215	1.627	2.076	2.944
Minimum	0.000	0.000	0.000	5.000	0.214	0.112	0.164	0.147
Maximum	4.000	6.000	4.000	6.000	1.429	1.739	2.240	3.091

3.2.4. Comparative study of notes in winter season

Fig. 18 and Fig. 19 show the comparative study of notes considering the day of week and value (kWh) respectively in the winter season. The maximum range recorded in the weekend dataset is 2.944 kWh due to more use of AC. Fig. 18, shows that the vacation is having maximum variance which is 4.044 and minimum for the weekend is 1.962.

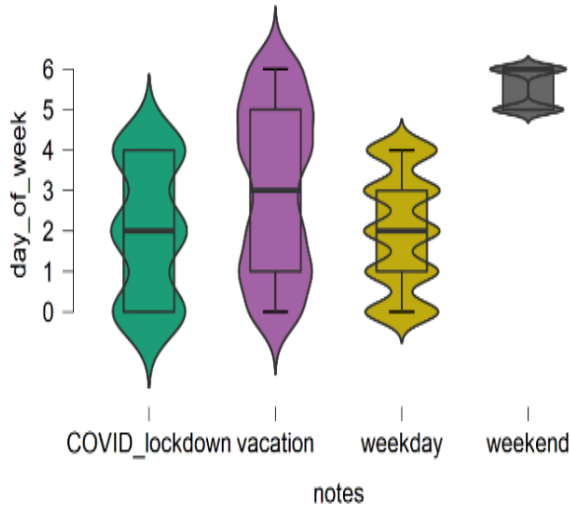


Figure 18. Comparison of notes with day of week

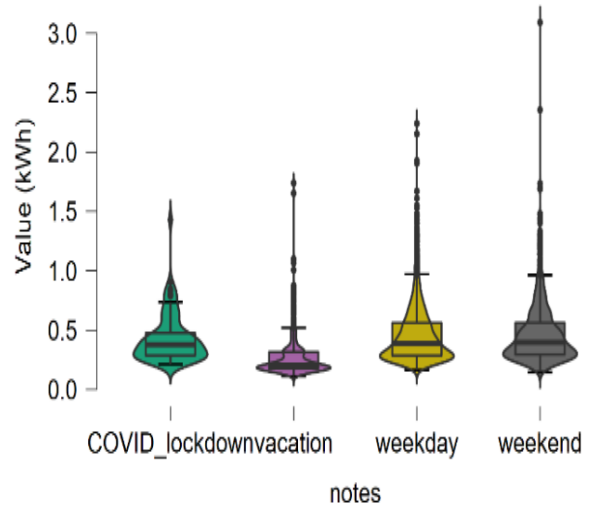


Figure 19. Comparison of notes with value (kWh)

4. 4. PREDICTION MODELS AND RESULT ANALYSIS

4.1 Orange Tool-based Predictive Modeling Method

Fig. 20, shows the workflow of the proposed predictive modeling. The workflow includes different parts like the training part, testing part, learning algorithm part, prediction and test-score evaluation part. In the next subsection proposed predictive modeling methodology has been discussed with actual values and finely tuned parameters for each algorithm and the results are obtained.

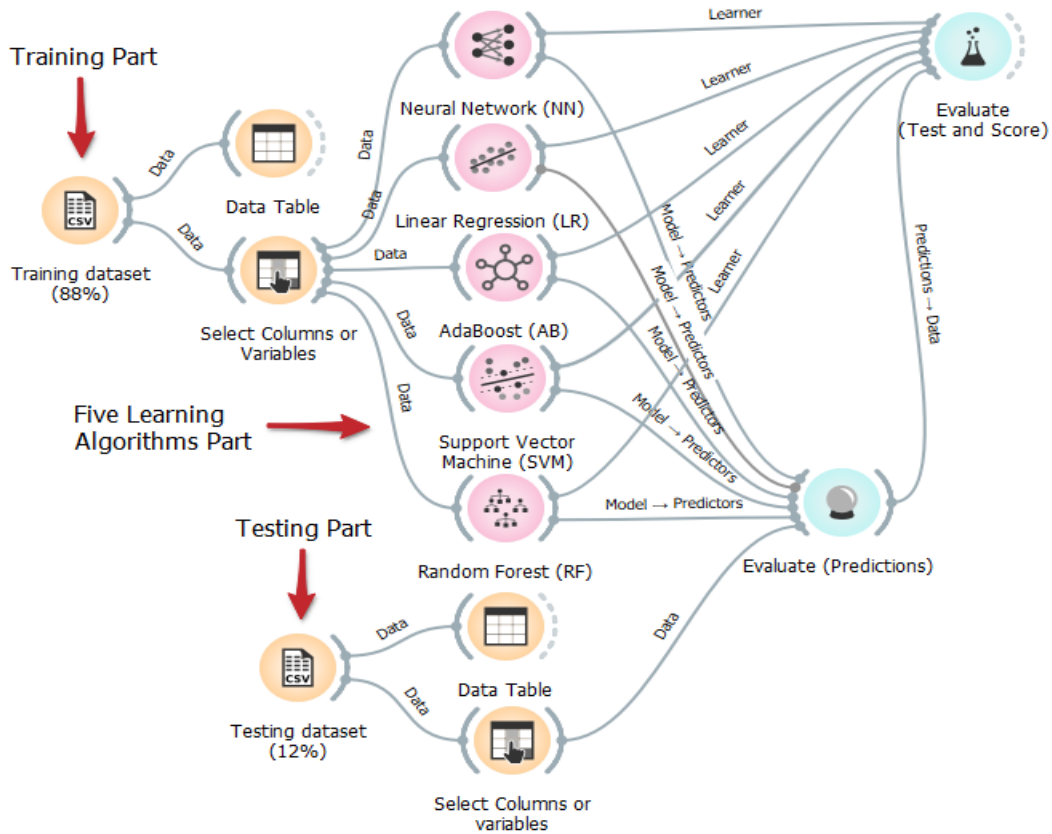


Figure 20. Workflow of proposed predictive modeling (<https://orangedatamining.com>)

Fig. 21 to Fig. 25 show the hyper parameters of different learning algorithms [12]. In the result analysis and discussion of the prediction model, the subsection enclosed the selected parameters with their fine-tuned values and obtained the results.

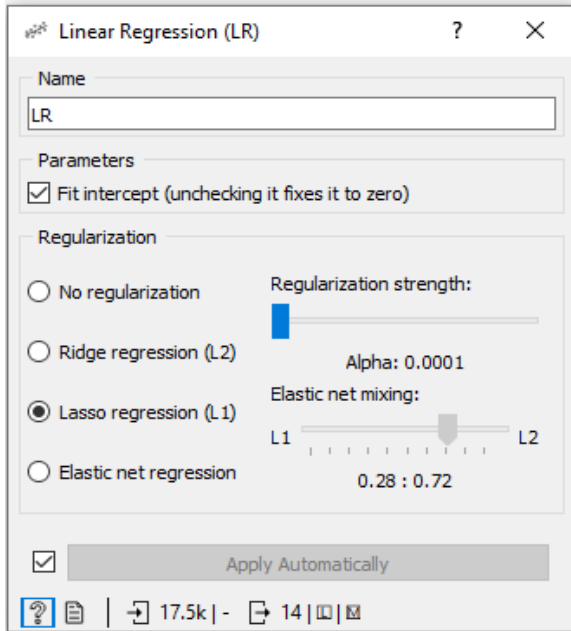


Figure 21. LR parameters

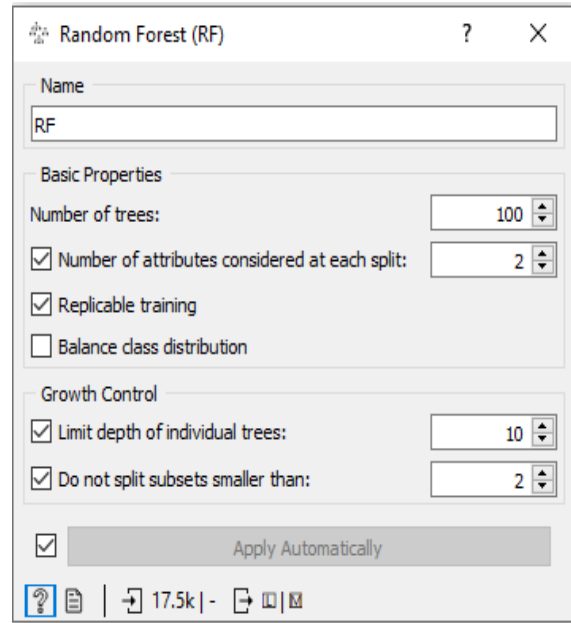


Figure 22. RF parameters

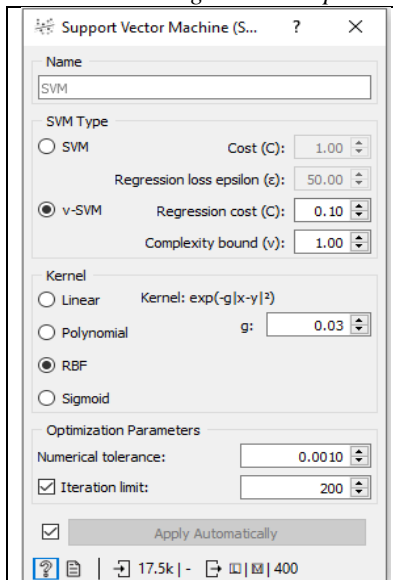


Figure 23. SVM parameters

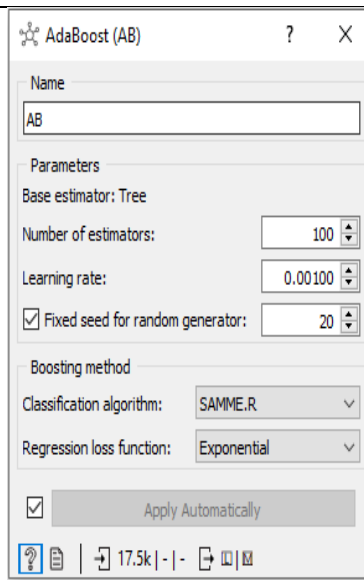


Figure 24. AB parameters

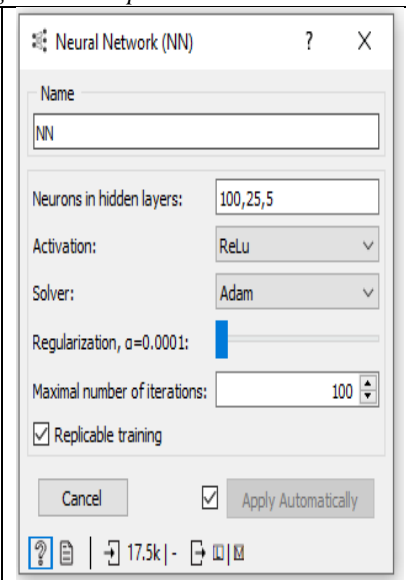


Figure 25. NN parameters

This subsection discusses the used five learning algorithms namely; LR, RF, SVM, AB and NN.

4.1.1. Linear regression model

The LR algorithm includes two input side features, the first is the input dataset and the second is the in-built preprocessor. The input dataset can be taken from the saved data of the Orange tool or can also provide externally from the proposed system. Moreover, the specialty of the Orange tool is that the algorithm-based widget is having an in-built preprocessing feature, which is automatically executing in a particular order. Under preprocessing the first step is to remove the row samples with the unknown

target values. The second step is to apply the one-hot-encoding process to categorical variables. The third step is to remove the column which is not having values. The fourth step is to replace the missing values by using the mean-based central tendency imputation method. The output side of the LR algorithm is having three important components namely: first is the linear regression learning algorithm which can learn the linear function from the input dataset, second is the trained model which can identify the associations between input predictor variables and output target variables. The third is the linear regression coefficients. In addition to this, the LR algorithm is having three regularization parameters namely: L1 (LASSO), L2 (ridge) and L1L2 (elastic net) as shown in Fig. 21. The LR algorithm used the first LASSO-based L1-norm penalty regularization parameter, which used can reduce the penalized version of the least square loss function. Similarly, the second is regularization parameters is ridge parameter used L2 penalty and third regularization parameters, which are the combination of L1 and L2 called as Elastic net regularization (orangedatamining.com/widget-catalog/model/linearregression/) [12].

4.1.2 Random forest model

The RF algorithm is having similar input and output features compared to the LR model except in RF the learner is the combination of decision trees called the Random Forest learning algorithm. In all trees, each tree is based on a bootstrap sample from the training data. Moreover, in the development phase of each tree the arbitrary subset of attributes is drawn randomly and from which better attribute can be selected for the split. From this process, the final RF model is developed based on the majority vote from individually developed trees. The RF model includes two main parameters namely: basic properties and growth control as shown in Fig. 21. Under basic parameters, it is needed to specify the number of trees as part of the forest. It is possible to choose the number of attributes which is required to draw randomly at each node.

If not the option is not selected or unchecked, then it will consider this number equal to the square root of the number of attributes in the data. The significance of the replicable training parameter is to obtain the replicability of results by fixing the seed for tree generation. The balance class distribution property is dealing with class balancing for the improvement of model performance. In the original work of the author [13] without controlling the growth of the tree, but after specifying the limit of the depth of the tree, the performance was improved. The RF model included one parameter to limit the depth of individual trees. Another growth control parameter is about to limit the split subsets. For the fitting, the model needs to apply the trial and error method and fine-tune the parameters. Moreover, the RF model is used for different applications like classification, regression and other tasks. This model was initially projected by Tin Kam Ho and then developed by authors Leo Breiman and Adele Cutler (<http://orangedatamining.com/widget-catalog/model/randomforest/>) [12, 13].

4.1.3. Support vector machine model

The SVM widget is used for both classification and regression types of problems. The specialty of the SVM model is easy to map the inputs with the higher-dimensional feature spaces. Moreover, the linear regression learning algorithm is the learner and instances are considered as support vectors. Based on the various minimization of the error function there are two types of SVM. The first SVM is the Epsilon-SVM model which applies to regression problems and another v-SVM model which applies to classification and regression type of problems as shown in Fig. 23. Moreover, the SVM is a Machine Learning (ML) method in which divides the attribute space with the hyperplane. This method is giving better predictive performance results by fine-tuning the different hyperparameters like: good setting of regression cost (C), regression loss epsilon (ϵ) and kernel (<http://orangedatamining.com/widget-catalog/model/svm/>) [12].

4.1.4. Adaptive boosting model

AdaBoost is the short form of Adaptive boosting ML algorithm which is used for both classification and regression problems. This algorithm is formulated by Yoav Freund and Robert Schapire. Moreover, it is a hybrid meta-algorithm that can be used to enhance the performance of weak learners. The learning algorithm is also available on the input side, which is not in the remaining four algorithms. On the output side, the AB learning algorithm is used as a learner. There are three main basic parameters like base estimator (tree), number of estimators and learning rate. Moreover, boosting methods are of two types based on the type of problems that is classification algorithms (SAMME and SAMME.R) and regression loss function (linear (), square (), and exponential ()) as shown in Fig. 24 (<https://orangedatamining.com/widget-catalog/model/adaboost/>) [12, 13].

4.1.5 Neural network model

The NN has the same input and output side features as discussed in the above LR model except in preprocessing consist of a normalization process using centering to mean and scaling to a standard deviation of 1. The NN model is having Sklearn based multi-layer perceptron learning algorithm with backpropagation. This NN is having the capability to learn both linear and non-linear models. There are a total of five main parameters namely: neurons per hidden layer, activation function for the hidden layer, solver for weight optimization, alpha (L2 penalty for regularization) and max number of iterations as shown in Fig. 25. There are four types of activation functions namely: identity (for implementing linear bottleneck), logistic function (logistic sigmoid function), tanh function (hyperbolic tanh) and ReLu function (rectified linear unit function). There are three weight optimizer solvers namely: L-BFGS-B (family of quasi-Newton methods), SGD (stochastic gradient descent) and Adam (stochastic gradient-based optimizer) (<https://orangedatamining.com/widget-catalog/model/neuralnetwork/>) [12].

4.2 Proposed Predictive Modeling Methodology

This paper uses a secondary data set, which is available on the Kaggle database. The database is in Comma-Separated Values (CSV) format. The dataset includes hourly energy consumption of two households located in Houston, Texas, USA. The Orange software is used to build five predictive models namely: Random Forest (RF), Linear Regression (LR), Support Vector Machine (SVM), Neural Network (NN) and Adaptive Boosting (AB) prediction models. There are two main aims of developing predictive modeling. First one is to validate the traditional predictive model performance under Covid-19 pandemic at HEC. The second one is to identify the most fitted predictive model. For the prediction analysis the whole dataset is split into training dataset with 17,520 hours (88%) and testing dataset with 2,305 hours (12%). For training, the model used the dataset from 1st January 2018 to 31st December 2019. Whereas, for testing model Covid-19 pandemic data of duration 4th January to 6th July 2020 is used. This model tries to determine the performance of traditional predictive models during anomalous situation like Covid-19 which can be used in the future for prediction. Moreover, a comparison of the predictive models using performance matrix namely: RF, LR, SVM, NN and AB prediction models has been done. The predictive modeling consists of three main stages as below:

Stage 1: To develop five predictive models at household energy consumption in Covid-19.

Stage 2: To select the best features, which will give better accuracy results.

Stage 3: To find the most fitted predictive model.

This methodology is again divided into four steps to perform energy consumption predictive modeling. Following are the steps followed to develop the predictive models [14]. Fig. 20 and Fig. 26 show the model building and testing flow in the Orange tool [12].

Following are the steps involved in developing the predictive model are mentioned by ML algorithms:

Step 1: Orange tool is having different models. Here five basic predictive models are selected to check its predictive capability on anomalies like Covid-19. The secondary dataset is used for prediction purposes. The source of the database from the open Kaggle repository.

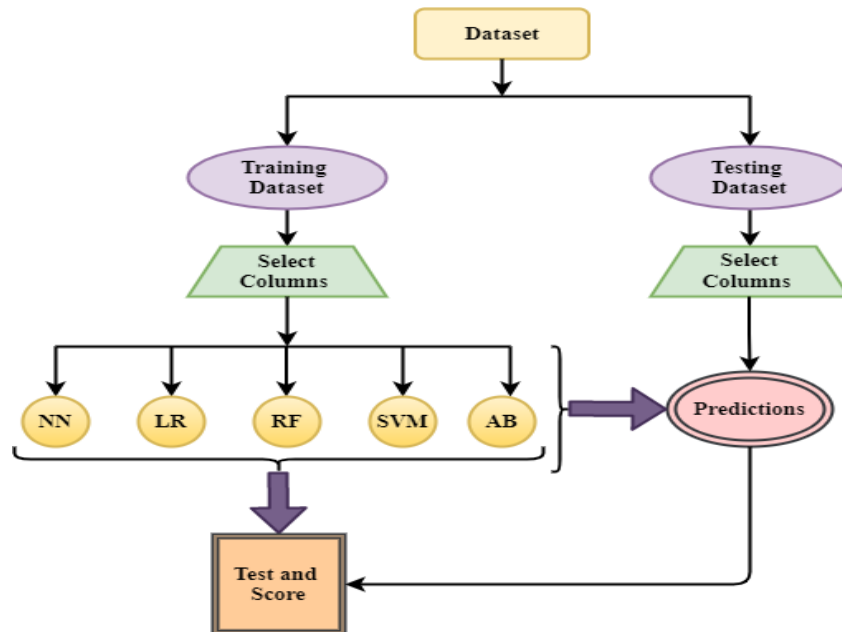


Figure 25. Overall workflow of prediction model

Step 2: In Orange software building predictive models including two sub-models one is by selecting the five models (LR, RF, SVM, NN and AB) as a training model and the second sub-model is to test the model on Covid-19 HEC dataset. And also adjusting the setting of each model parameter to fit the model and get better accuracy.

Step 3. To select the more accurate and well fitted predictive model to use further for predicting the HEC during anomaly like Covid-19. The parameters in the performance matrix consist of few parameters namely: MSE, RMSE, and Mean Absolute Error MAE [14].

4.3 Result Analysis and Discussion of Prediction Models

The Orange tool is used for the development of predictive modeling, testing and score analysing [12]. Fig. 26 gives the comparative analysis of five prediction models. These results are achieved after tuning the parameters of each model. From the five considered model, NN model shows better predictive performance then other models as shown in Fig. 27.

The tuned NN model has three hidden layers, Relu activation function and Adam solver with a maximal number of 100 iterations. The LR model uses Lasso regression (L1) for regularization of the model. The AB model used a tree as a base estimator with 100 estimators. The learning rate of AB is fixed at 0.00100 with 20 as a fixed seed for the random generator. The AB model includes two boosting methods based on the classification algorithm and regression loss function.

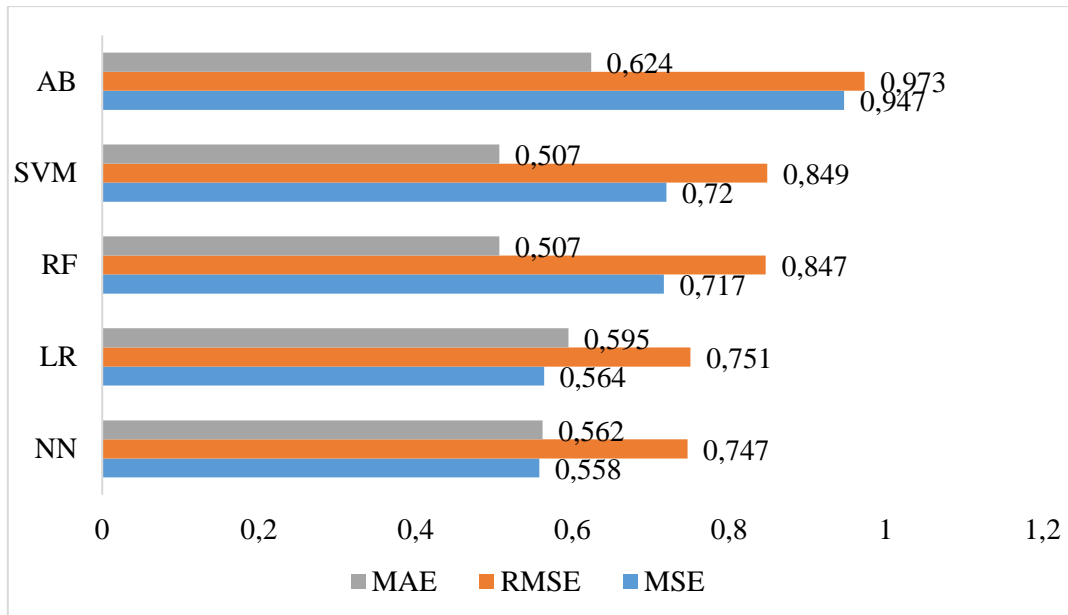


Figure 26. Prediction result of five models

The prediction analysis achieved better results with SAMME. R classification algorithm and exponential regression function. In SVM model, the complexity bound ν -SVM type gives better prediction results compared to other types of SVM model. The regression cost (C) is fixed to 0.10 and complexity bound (ν) is fixed to 1 value. The Radial Basis Function (RBF) is selected as Kernel. The two optimization parameters, tolerance and iteration limit is set to 0.0010 and 200 respectively. For the RF model the basic parameters are set as follows: The number of trees is adjusted to 100, number of attributes considered at each split is 2 and replicable training parameter is selected. The growth control parameters like the limit depth of individual trees is set to 10 and not split sub-sets value is taken less than 2. The parameters of the performance matrix are MSE, RMSE and MAE.

Table 4 to Table 6 show the test and score results based on sampling methods. The RF model shows better test and score results than other models as shown in Table 4 and Table 5. Table 4, considered the stratified, 20-fold cross-validation sampling method. Table 5, results are based on random sampling method including stratified shuffle split with repeat train/test value of 10 and 66% training set size. The AB model gives better performance on the train data sampling method as shown in Table 6.

Table 4. Stratified 20-fold cross validation

Model	MSE	RMSE	MAE
NN	0.634	0.796	0.504
LR	0.458	0.676	0.494
RF	0.231	0.481	0.302
SVM	0.745	0.863	0.555
AB	0.288	0.537	0.301

Table 5. Stratified shuffle split

Model	MSE	RMSE	MAE
NN	0.650	0.806	0.509
LR	0.466	0.683	0.493
RF	0.262	0.512	0.321
SVM	0.746	0.863	0.587
AB	0.313	0.559	0.314

Table 6. Test on training data

Model	MSE	RMSE	MAE
NN	0.634	0.796	0.504
LR	0.455	0.674	0.493
RF	0.114	0.338	0.222
SVM	0.654	0.808	0.487
AB	0.000	0.002	0.000

5. CONCLUSION AND FUTURE SCOPE

In the present scenario integration of distributed generation with the power grid and energy demand is increasing day by day. So, demand-side management is a key element to achieve effective energy planning and energy optimization on the distribution side. For effective and micro-level planning the demand of the individual consumer should be known prior. The same analysis is very challenging in the residential sector due to the diverse and unpredicted use of load. The usage is mainly influenced by the end-user behaviour. This is majorly influenced by calendar and seasonal factors. This paper applied the statistical and predictive modeling technique to understand the performance of traditional predictive models on household electricity consumption during anomalies like Covid-19. The effects of a calendar and seasonal factors on household energy consumption are considered. For the analysis, the Kaggle database is used which is of the labeled type and time series in nature. The detailed analysis is done based on calendar periods such as Covid-lockdown, vacation, weekday, and weekend. From the predictive modeling study, the NN model is the most fitted model than other models followed by the LR model. The values of the performance matrix parameter like MSE, RMSE and MAE of the NN model is observed to be 0.558, 0.747 and 0.562 respectively. This study would benefit the researchers and utility companies for prediction of household electricity consumption under anomalous situations like Covid-19. Moreover, the usage of Orange and JASP for statistical and predictive modeling analysis has been discussed.

As a future scope, the analysis could be extended for the effects of the calendar and seasonal variability on the used individual appliances. The effects of a calendar and seasonal dataset on the accuracy of the prediction model can be analysed.

Acknowledgment

The author would like to thank Symbiosis International (Deemed University) for permitting to carry out the proposed research and to use resources to accomplish the objectives.

REFERENCES

- [1] Wen, L, Zhou, K, Yang, S. Load demand forecasting of residential buildings using a deep learning model. *Electric Power Systems Research* 2020; 179: 106073, DOI: 10.1016/j.epsr.2019.106073
- [2] Teeraratkul, T, O'Neill, D, Lall, S. Shape-Based Approach to Household Electric Load Curve Clustering and Prediction. *IEEE Transactions on Smart Grid* 2017; 9(5): 5196–206, DOI: 10.1109/TSG.2017.2683461
- [3] Moezzi, M, Lutzenhiser, L. What's Missing in Theories of the Residential Energy User. 2010 *ACEEE Summer Study on Energy Efficiency in Buildings, Center for Urban Studies Publications and Reports* 2010; 207–21. <http://archives.pdx.edu/ds/psu/22747>
- [4] Amasyali, K, El-Gohary, NM. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews* 2017; 81: 1192–205. DOI:10.1016/j.rser.2017.04.095
- [5] Chunekar, A, Sreenivas, A. Towards an understanding of residential electricity consumption in India. *Building Research and Information* 2018; 47(1): 75–90, DOI: 10.1080/09613218.2018.1489476
- [6] Li, C, Song, Y, Kaza, N. Urban form and household electricity consumption: A multilevel study. *Energy and Buildings* 2018; 158: 181–93, DOI: 10.1016/j.enbuild.2017.10.007
- [7] Lusi, P, Khalilpour, KR, Andrew L, Liebman A. Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Applied Energy* 2017; 205: 654–69. DOI:10.1016/j.apenergy.2017.07.114
- [8] Chitsaz, H, Shaker, H, Zareipour, H, Wood D, Amjady N. Short-term electricity load forecasting of buildings in microgrids. *Energy and Buildings* 2015; 99: 50–60, DOI: 10.1016/j.enbuild.2015.04.011
- [9] Hippert, HS, Pedreira, CE, Souza, RC. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems* 2001; 16(1): 44–55, DOI: 10.1109/59.910780
- [10] Li, M, Allinson, D, He, M. Seasonal variation in household electricity demand: A comparison of monitored and synthetic daily load profiles. *Energy and Buildings* 2018; 179:292–300, DOI: 10.1016/j.enbuild.2018.09.018
- [11] Love, J, Selker, R, Marsman, M, Jamil, T, Dropmann, D, Verhagen, J, Ly, A, Gronau, QF, Šmíra, M,

- Epskamp, S, Matzke, D, Wild, A, Knight, P, Rouder, JN, Morey, RD, Wagenmakers, E. JASP : Graphical statistical software for common statistical designs. *Journal of Statistical Software* 2019; 88 (2), DOI: 10.18637/jss.v088.i02
- [12] Demšar J, Curk, T, Erjavec, A, Gorup, C, Hocevar, T, Milutinovic, M, Mozina, M, Polajnar, M, Toplak, M, Staric, A, Stajdohar, M, Umek, L, Zagar, L, Zbontar, J, Zitnik, M, Zupan, B. Orange: data mining toolbox in python. *Journal of Machine Learning Research* 2013; 14(1): 2349-2353
- [13] Breiman, L. Random forests. *Machine learning* 2001; 45(1): 5-32.
- [14] Solyali D. A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus. *Sustainability (Switzerland)* 2020; 12(9), DOI: 10.3390/SU12093612