MEF UNIVERSITY

DEVELOPMENT AND COMPARISON OF PREDICTION MODELS FOR ESTIMATING SHORT TERM ENERGY DEMAND OF A HOTEL BUILDING

Capstone Project

Selimcan Yılmaz

İSTANBUL, 2017

MEF UNIVERSITY

DEVELOPMENT AND COMPARISON OF PREDICTION MODELS FOR ESTIMATING SHORT TERM ENERGY DEMAND OF A HOTEL BUILDING

Capstone Project

Selimcan Yılmaz

Advisor: Prof. Dr. Özgür Özlük

İSTANBUL, 2017

EXECUTIVE SUMMARY

DEVELOPMENT AND COMPARISON OF
PREDICTION MODELS FOR ESTIMATING
SHORT TERM ENERGY DEMAND
OF A HOTEL BUILDING

Selimcan Yılmaz

Advisor: Prof. Dr. Özgür Özlük

SEPTEMBER, 2017, 15 pages

This project presents a machine learning model building approach to developing a model for predicting next hour electricity consumption of a hotel complex in Cyprus, with the aim of improving existing prediction accuracy due to comparing different models to choose best performing. Model building process in this project includes three main steps. First data understanding and processing, second; feature extraction and selection, model building and implementation, third; compare of results that obtained by different models.

For this project to build a successful model, I used the real world data that provided by the company "ReEngen"; which operates in energy efficiency and provides its customers a platform as a service using IOT devices to collect energy consumption data and interpret. Company collaborates with MEF University in several topics and provided a large dataset that collected from one of their clients. Main objective is deploying a model that considers previous inputs and forecasts the value of next hour's possible energy consumption of a large hotel complex in Cyprus which actively operates 24 hours in a day and 12 months in a year. Main objective of first phase was to step up on existing model's performance which is around 70%. I choose two of popular machine learning algorithms which are artificial neural network and random forest to deploy a prediction model. Prediction accuracies of the models are 90% and 89% in order, which are obviously higher than existing models that company is using. Results show that previous examples and data can be used to predict future electricity demand values in some aspects.

Key Words: electricity demand forecast, neural network, random forest, demand prediction, time series analysis.

ÖZET

BİR OTEL KOMPLEKSİ İÇİN
KISA VADELİ ELEKTRİK TÜKETİMİ
TAHMİN MODELLERİ OLUŞTURULMASI VE
MODELLERİN KARŞILAŞTIRILMASI

Selimcan Yılmaz

Tez Danışmanı: Prof. Dr. Özgür Özlük

EYLÜL, 2017, 15 Sayfa

Enerji verimliliği alanında çalışan ve müşterilerine enerji tüketim yönetimi yapabilecekleri bir yazılım ürünü sağlayan "ReEngen" firması ile MEF Üniversitesi'nin işbirliği, bu projenin ortaya çıkmasına yardımcı olmuştur. Enerji verimliliği günümüzün revaçta konularından bir tanesidir. Gezegenimizin karşı karşıya bulunduğu küresel ısınma gibi ciddi meseleler; enerji verimliliği ve akıllı enerji yönetimi konularının önemini fevkalade artırmıştır.

Enerji verimliliğini yönetebilmenin bir yolu da iyi bir tahmin modeli kurabilmektir.. Bu projede yapmaya çalıştığımız konu da, geçmiş tüketim verilerini inceleyerek bir sonraki periyodun enerji tüketimini yüksek bir keskinlikte doğru tahmin edebilmektir. Projeye konu olan zaman aralığı saatliktir. Çalışma sonucunda firmanın mevcut tahmin modellerinin (%70) üzerine skor üreten, yapay sinir ağları (%90) ve rastgele orman (%89) algoritmalarını kullanan 2 adet tahmin modeli geliştirilmiştir.

Anahtar Kelimeler: enerji talep tahmini, yapay sinir ağları, rastgele orman, karar ağaçları, talep tahmini, zaman serisi analizi

TABLE OF CONTENTS

1. INTRODUCTION

Forecasting the future has always been an important topic for history of humanity. Uncertainty and unknowable structure of future makes the concept of prediction mysterious and attractive. Prediction and forecast techniques are always been a topic of interest for researchers and scientists. Forecast in time series data also takes up space in mostly economics and finance, also in energy demand analysis increasingly. The aim of this project is to execute a successful prediction model that forecasts short term electricity demand of a hotel complex in Cyprus.

When it comes to electricity consumption, demand planning and operating an electricity distribution system is very important because of non-storable structure of electricity. Therefore, provider companies of electricity have to continually provide users with electricity in adequate volumes at reasonable voltage and ensure a reliable electricity distribution system while considering the non-storability of electricity. Short term and long term energy demand forecast and estimation of future usage is essential for managing and operating an electrical supply and distribution system efficiently. Demand prediction of electricity is important for activities, such as, new system developments and expansions, forecasting operational costs of new sub-distribution systems, transformers, and adequateness of infrastructure capacities. Dynamic pricing and demand management are also important. At the beginning of this project, the main object was to evaluate a data set of a distribution substation, however, due to restrictions and delays, the success of this project was unlikely. Therefore, after negotiating with the data provider company, we decided to use the data set of a commercial hotel complex and considered building energy efficiency topic as the subject. To achieve this, first, a preliminary analysis was conducted with the transformer data. Second, the outcomes and methodology of the preliminary analysis were used to analyze the data from the hotel complex in order to achieve our project objective, which is setting a short term energy demand prediction model.

The data is provided from the company named "ReEngen". They provide building IoT solutions. The company has a building energy management platform which uses the data collected from the building with IoT devices. Through data analytics and optimization technologies in coordination with smart sensors, controls and meters; company creates a track record of increased efficiency in energy management market.

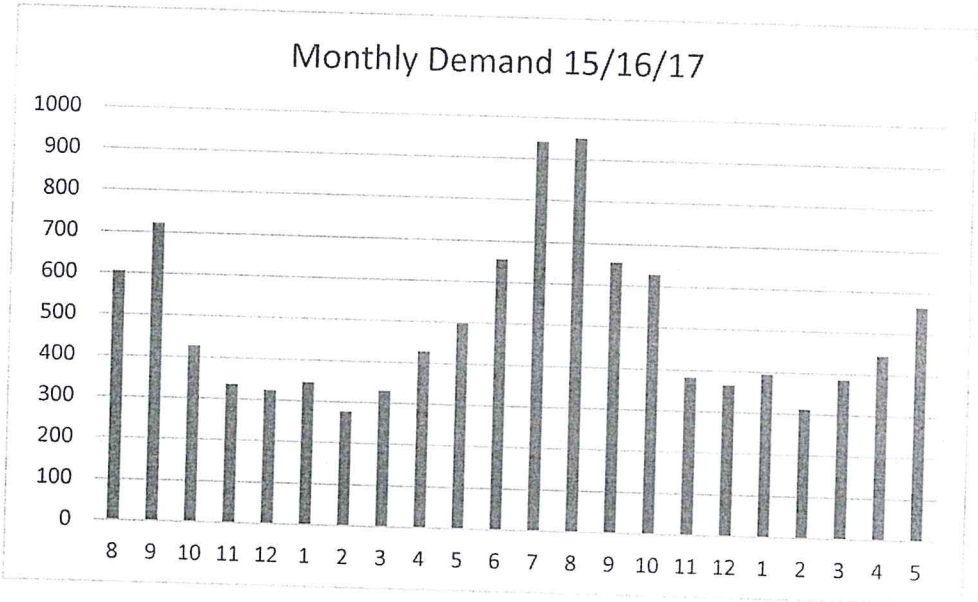
Researchers interested in energy consumption prediction and tried to use different methods and techniques to improve building energy efficiency and performance. The short-term predictions mainly focus on the forecast of the daily energy demand, peaks, and loading profiles [1]. Approaches for developing prediction models include certain linear, machine learning and artificial intelligence methods [2-6]. Artificial neural networks and random forests are also implemented on time series data for predicting energy demand forecast in various studies. [7-10].

2. ABOUT THE DATA

The time series data provided by the company "ReEngen" contains 14.388 observations, starting from August 2015 and ending at June 2017. There are four columns in the data set, namely, the date, the hour, the energy used by the end of that hour (in terawatt/hour-twh), and measured temperature during that hour (in C).

Looking at the monthly period, a certain seasonality that can be observed easily in the data set (see Figure 1).

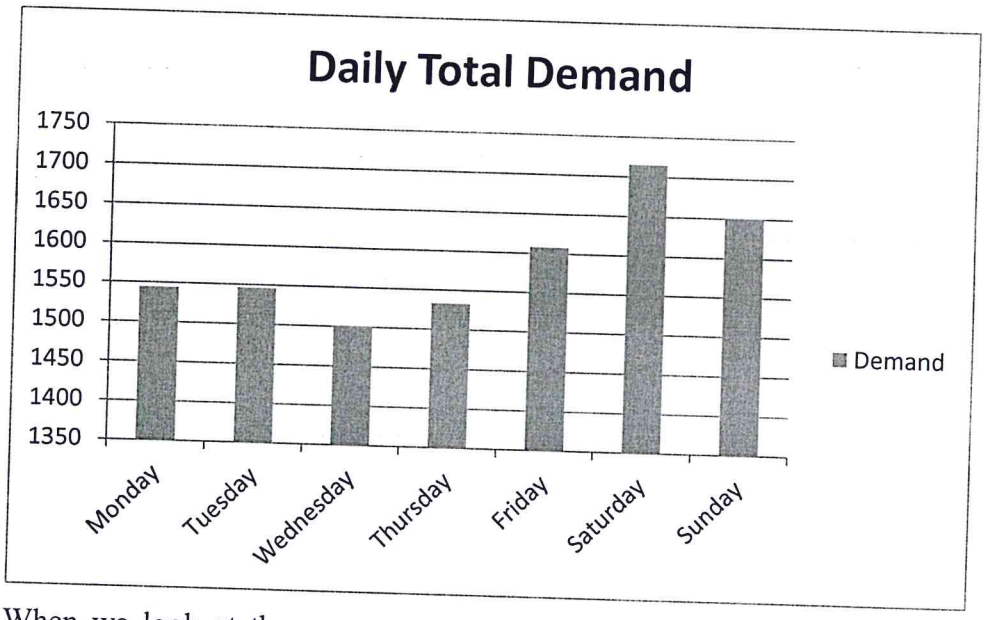
Figure 1 The Monthly Frequency of the Electricity Demand in the Hotel Complex



This was a predicted outcome when we considered with domain knowledge, because this hotel, located in Cyprus, is a sea side hotel which means it has more activity in the summer season and activities increase electricity consumption. In the exploratory analysis, a strong positive correlation was detected between electric demand and the temperature (r=.76).

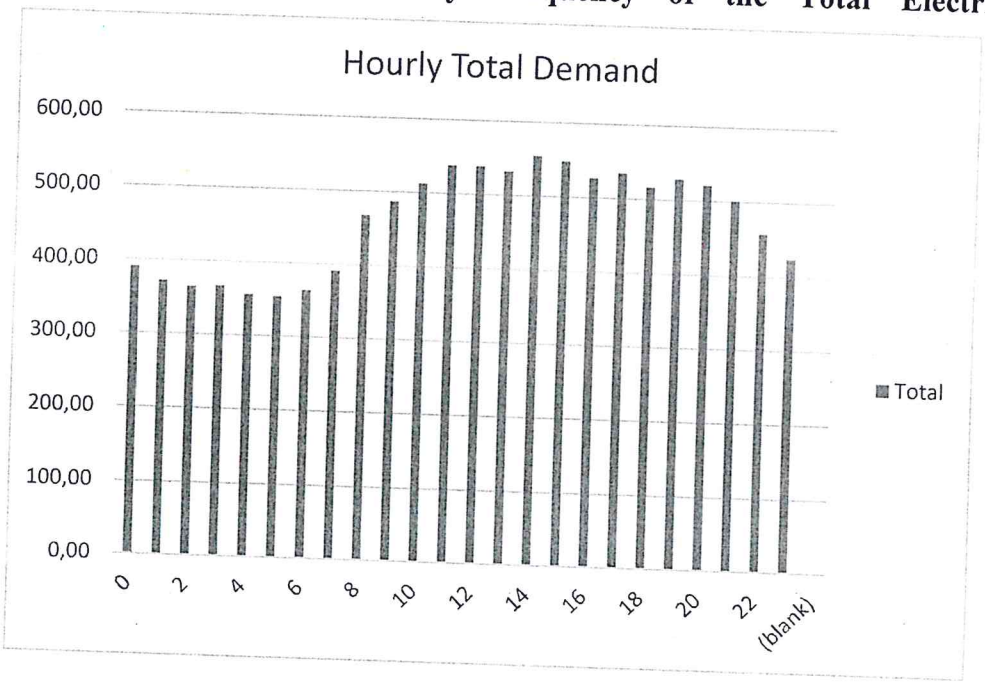
Additionally, at the daily scale, general energy demand of the hotel remains almost the same during the whole week and doesn't seem to change depending on the day of the week. However, on Saturdays, electricity consumption seems to be a little higher than other days of the week (see Figure 2).

Figure 2 The Daily Frequency of the Total Electricity Demand in the Hotel Complex



When we look at the energy consumption at the hour of the day, there are two visibly different periods. The first one starts at 8am, with beginning of the day and ends with 11pm, with the end of the night. Second shift starts with 11pm and end in the morning before 8am. Considering the shifts and active hours of a hotel and its customers this was an expected outcome. (see Figure 3)

Figure 3 The Hourly Frequency of the Total Electricity Demand



To conclude, electricity usage of hotel continues whole year by changing scale. At season of summer electricity consumption increases, at Saturdays consumption is higher than other days as expected, and days have 2 shifts within.

2.1. Missing Value Handling and Outlier Detection & Handling

The data set provided by this company was generally clean and steady, however, there were missing data and huge deviations occurred at some data points and these enormous deviations from mean made statistical analysis very hard to implement. Therefore, data cleaning and preparation played a huge role in this project.

The company uses IoT devices and some modern techniques to collect energy consumption data from utility systems of buildings and outside temperature. However, these devices may have some breakdowns and some malfunctioning during the collection phase. These problems negatively affect the data collection process. There were 2 common problems observed about the data quality; which were missing values and outliers.

First problem observed was missing values. Around 1500 hour long data points were missing because devices were unable to collect data at that time. Dealing with missing values is an important step in model building. I used "interpolate" function of "pandas" library in python. This function examines linearity of data and fills missing points accordingly.

Second one is outliers that occurred because of overload on a data point. In this case, when the device could not record data for a certain time (for example, leaves 6 hours empty) it may then record the next hour's input as the sum of the previous data points. This situation creates strong deviations at our time series and leads to the observation of the outliers. To deal with outliers, I used the method that company has already been using which is looking at how many hours are missing before outlier data point, then dividing the value to the number of missing hours, and imputation of calculated values to the missing points.

2.2. Checking If The Data Is Stationary Or Not

In the exploratory data analysis, the data set was observed to behave linearly and tended to oscillate around the mean. To understand the data better, after handled missing values and outliers, an Augmented Dickey-Fuller test (ADF) [11] was conducted to check whether the data is stationary or not (if not, this may mean that the data has unit root and it

increases or decreases over time in a trend). As a result, enough statistical evidence was found to reject the null hypothesis (data is not stationary), and the data may be considered as stationary at the 95% confidence interval (see Table 1).

Table 1 The Result of the ADF on the electricity demand

```
ADF Statistic: -3.087217
p-value: 0.027514
Critical Values:
        1%: -3.431
        5%: -2.862
        10%: -2.567
```

The result of the ADF test can be considered as a proof of the importance of following further analysis, because it was essential to understand the data clearly. The data being stationary means that auto regressive analysis may be done and it allows the user to employ previous input to predict following data point [12].

3. PROJECT DEFINITION

3.1. Problem Statement

The company wanted to increase accuracies of existing prediction models with the help of machine learning methods to make better forecast, planning and energy consumption management.

Today, traditional prediction methods based on linear or time series models that company is using could not pass over 70% accuracy scores. The company wanted to go beyond existing models which they have been using and try non-parametric regression methods. These methods also will help finding non-linear relationships that cannot be found with linear models in the data set. Solving this problem will improve prediction strength of company's energy management platform and support energy efficiency management process.

To do this, artificial neural network and random forest algorithms were chosen for execution of this project and results are compared with each other.

3.2. Project Objectives

First objective of this project was going beyond existing accuracy levels (70%) of the models that company has already been using. Second objective was choosing a non-parametric prediction algorithm that will approach to the problem differently than existing linear models. Third and last objective was setting a proper, simple and efficient prediction model that works fast and capable of making instant calculations quickly for hourly periods.

3.3. Project Scope

A model that works for hourly periods was satisfactory at the first step. Building a model to make daily forecast was not a suitable goal, because data quantity is not sufficient to establish a healthy model that works for daily periods (only had 565 days). Models that take 5 and 15- minute scales as input were not expected by the company. Project scope was also limited to the prediction of electricity consumption because of lack of data. Prediction of temperature values were also out of the project scope.

4. METHODOLOGY

4.1. Tools

In this project, python and its certain well known libraries were used during model building process. "Scikit-learn" library was used for machine learning operations and algorithm applications. "Numpy" and "Pandas" libraries were used for data pre-processing and data manipulation. "Statsmodels" was used to implement the ADF test. MS Excel was also used for data cleaning and visualization.

4.2. Data Requirements and Sources

After the missing values were cleaned and outlier effects were eliminated, the next step was the feature extraction. Since the aim of this project was to analyze previous values to predict next value, which previous values the model will take as an input should be decided. To resolve this, auto correlation table of both electricity demand and temperature were used and last hour's correlation between previous hours was examined (see Figures 5 and 6).

Figure 5

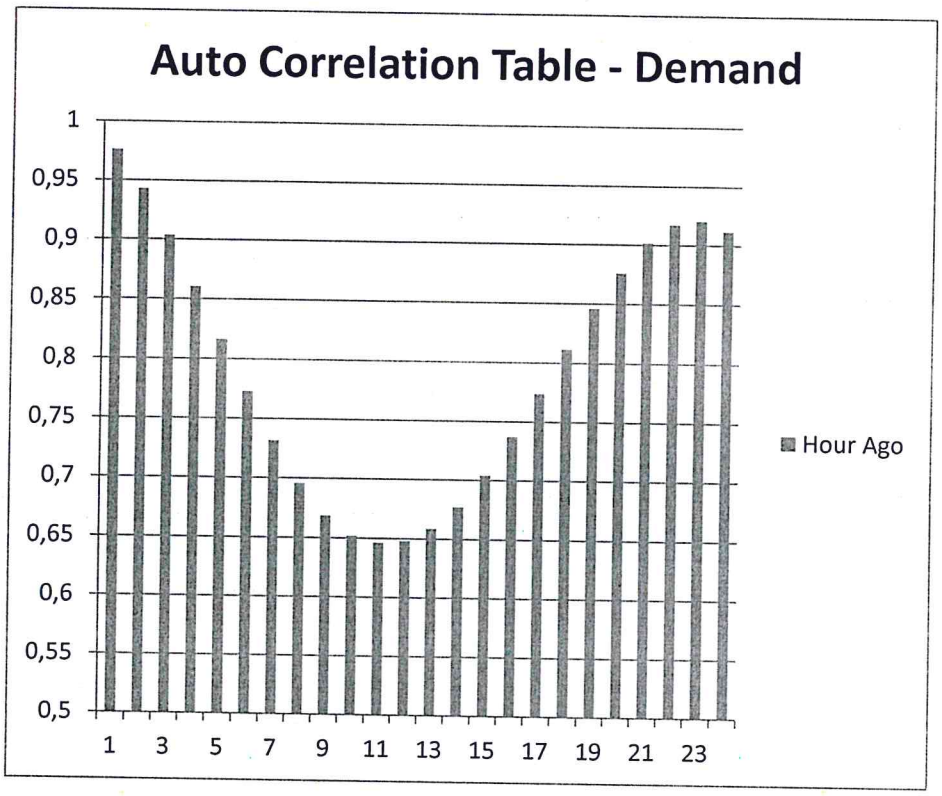
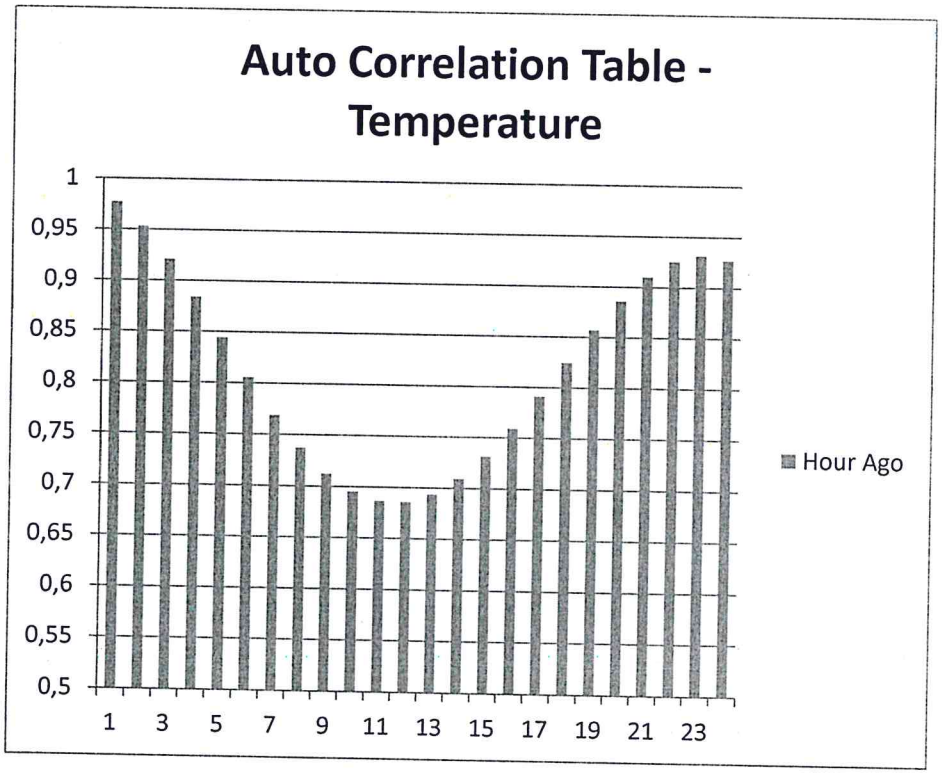


Figure 6



Lots of combinations were attempted and used as input in the models, eventually up to previous 3 hours of consumption and temperature values fulfilled the need and they were selected as input. Using more lags and more inputs increased R-Squared of training sets, but did not have a significant impact on test and validation sets.

Additionally, checking if the day is weekday (1 for weekend, 0 for weekend) and which shift belongs to the related hour during that time (1 for morning shift, 0 for night shift) were added as input.

Eventually, data set increased to 8 columns and 1 target column to predict the energy consumption at that hour. Eight columns consist of 3 lags for energy consumption, 3 lags for prior temperature, 1 weekday-weekend difference, and 1 shift difference.

4.3. Methods and Techniques

After new features are extracted and the data set came to final state, the data was split into 2 main parts which are train-test part and validation part. Data until the end of 2016 used for model training, testing and the model development process. Data belongs to 2017 which are 3.744 rows (26% of total) allocated as validation data and excluded from train-test part (74% - 10.643 rows).

Following these operations, two algorithms were chosen, namely, neural network and random forest algorithms. These models were deployed and fitted to the train & test data. 70% of the data was used to train the model and 30% of the data was used for testing.

After the splitting step, fundamental parameters were optimized in algorithms. For artificial neural network, "relu" function used for activation (the rectified linear unit function returns $f(x) = \max(0, x)$) and for solving weight optimization, stochastic gradient based optimizer "adam" parameter is used. Two hidden layers, sequentially with 150 and 50 neurons, were used for solving the problem. In random forest algorithm, for "n_estimators" parameter, which controls the number of classifying decision trees on data set, value 250 was used. For the rest of both algorithms default parameters were used (see Appendix 1).

After 5 fold cross validation for both train and test sets, models were applied to the validation data set and performances were monitored. R-squared and root mean squared error (RMSE) metrics were used for model performance comparison.

5. RESULTS

Results show that (see Table 2), for both training and test data, Random Forest model performs better at model building stage, where the accuracy - r squared of RF algorithm is higher and root mean squared error (RMSE) is lower than Artificial Neural Network.

Table 2 Model Comparisons

MODELS	TRAIN R2	TEST R2	TEST RMSE	VALIDATION R2	VALIDATION RMSE
RF	**0,96**	**0,96**	**0,070**	0,894	0,061
ANN	0,94	0,92	0,081	**0,899**	**0,060**

However, for validation data set, ANN algorithm performs better than RF and RMSE of ANN is lower than RF algorithm. It is predicable that for this data set, although RF algorithm provided higher accuracy during training and test stage, ANN algorithm worked better at the data which the model was not previously exposed to. It shows that ANN actually learned patterns and nonlinear relationships better for this demand prediction regression problem.

In addition, these findings indicated that obtained results (89%) exceeded existing model performance (70%), as expected.

6. SOCIAL AND ETHICAL ASPECTS

When we look at the potential impact of the project on social, ethical and environmental levels; setting a proper and well working prediction system will help all stakeholders of the company. Both clients who are responsible from building utility systems and client's account managers in the company will benefit from this system. With this model they will be able to make better planning for monthly energy consumption of the building, and while monitoring daily and hourly energy demand, they will have an opportunity to evaluate monthly actual and expected results before the end of the month. This will lead better planning and management for the company which in the long run will enhance the efficient use of energy.

7. VALUE DELIVERED (CONTRIBUTION)

After our meetings with the company, the main value delivered was their genuine consideration to add this prediction model into their energy management platform. For its customers and stakeholders, the company which delivered the data of the project will bring this short term energy prediction model into use as a new feature of their energy management platform and add this model into their software. With this new feature, customers of the company will be able to plan and monitor their energy consumption easily. They will have a chance to take actions before consumptions and costs are realized, even before the end of the month.

As a result of this model building project, there will be a definite improvement in the company's prediction ability as the accuracy of models used increased to 89% from 70%. Adding this project's results into the energy management platform of the company will enhance the functionality of the platform.

Additionally, this feature will help the company to detect outliers and determine problems with the IoT machines that they are using for data collection from utility systems with comparing actual consumption and results of the model. For example, if the electricity consumption in an hour is or closer to 0, but the model prediction was much more greater than the number shown, then this deviation between the predicted and the actual result will indicate that there may be a difficulty with machines' data collection from the system.

In conclusion, the objective of this project was met and the company was provided with new prediction models with different algorithms than the existing which increased the accuracy, and eventually will lead to increase in the efficiency of their energy management.

APPENDIX 1

The python code of artificial neural network algorithm that established a model to predict short term energy demand.

```
from sklearn.neural_network import MLPRegressor
mlp_reg = MLPRegressor(hidden_layer_sizes=(150,50), activation='relu',
                       solver='adam', alpha=0.0001,
                       batch_size='auto', learning_rate='constant',
                       learning_rate_init=0.001, power_t=0.5,
                       max_iter=200, shuffle=True,
                       random_state=None, tol=0.0001,
                       verbose=False, warm_start=False,
                       momentum=0.9, nesterovs_momentum=True,
                       early_stopping=False, validation_fraction=0.1,
                       beta_1=0.9, beta_2=0.999, epsilon=1e-08)
```

The python code of random forest algorithm that established a model to predict short term energy demand.

```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor(n_estimators=250, criterion='mse',
                       max_depth=None, min_samples_split=2,
                       min_samples_leaf=1, min_weight_fraction_leaf=0.0,
                       max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       bootstrap=True, oob_score=False,
                       n_jobs=1, random_state=None,
                       verbose=0, warm_start=False)
```

REFERENCES

[1] Abdel-Aal, R. E. (2006). Modeling and forecasting electric daily peak loads using abductive networks. International Journal of Electrical Power & Energy Systems, 28(2), 133-141.

[2] Moazzami, M., Khodabakhshian, A., & Hooshmand, R. (2013). A new hybrid day-ahead peak load forecasting method for Iran's National Grid. Applied Energy, 101, 489-501.

[3] Yalcintas, M., & Akkurt, S. (2005). Artificial neural networks applications in building energy predictions and a case study for tropical climates. International journal of energy research, 29(10), 891-901.

[4] Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5), 545-553.

[5] Solomon, D., Winter, R. L., Boulanger, A. G., Anderson, R. N., & Wu, L. L. (2011). Forecasting energy demand in large commercial buildings using support vector machine regression. Department of Computer Science, Columbia University, Tech. Rep. CUCS-040-11.

[6] Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. IEEE transactions on power systems, 18(3), 1014-1020.

[7] Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy, 127, 1-10.

REFERENCES

[8] Foucquier, A., Robert, S., Suard, F., Stéphan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. Renewable and Sustainable Energy Reviews, 23, 272-288.

[9] Azadeh, A., Ghaderi, S. F., Tarverdian, S., & Saberi, M. (2007). Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption. Applied Mathematics and Computation, 186(2), 1731-1741.

[10]Qiu, X., Zhang, L., Ren, Y., Suganthan, P. N., & Amaratunga, G. (2014, December). Ensemble deep learning for regression and time series forecasting. In Computational Intelligence in Ensemble Learning (CIEL), 2014 IEEE Symposium on (pp. 1-6). IEEE.

[11]Cheung, Y. W., & Lai, K. S. (1995). Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business & Economic Statistics, 13(3), 277-280.

[12]Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.