भारतीय प्रौद्योगिकी संस्थान दिल्ली
Indian Institute of Technology Delhi

CLL788: Process Data Analytics

Instructor: Prof. Manojkumar Ramteke
Author: Utkarsh Dogra (2020CH70199)

Efficient Energy Consumption Prediction Model

Contents:

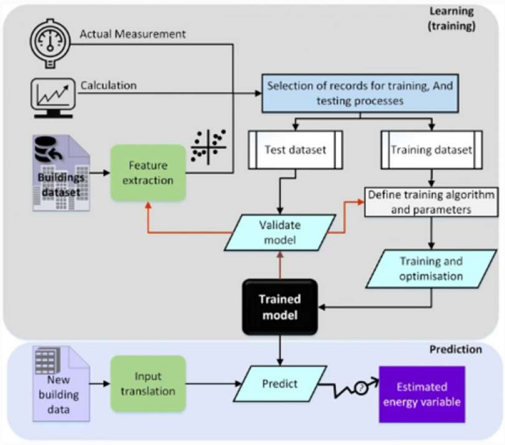

Fig. 1[4]: Predicting and Training Energy Consumption[1]

Abstract:

The ongoing development of the modern technology has increased the usage of energy. This makes it important to use the energy efficiently, and somehow know the expected energy demand at a particular time. Based on a known data from the Korea Electric Power Corporation, this report will try to find the type of load (whether Light, Medium or Maximum Load) based on certain known parameters.

Introduction:

The governments of different places are trying to adopt the concept of smart cities, and based on the datasets which are collected, it's tried to predict what is the possible energy requirement based on the time of a day, using the parameters like reactive power, time of the day, status of the day (weekday or the weekend) etc.

This report focuses on the prediction of the load type. The methods used for prediction are same as the methods used in the paper, basically making use of three different techniques, after preprocessing the data, which are[2]:

1) Linear Regression
2) SVM
3) Random Forest

The predictions are then evaluated based on Confusion Matrix. The results obtained from the Random Forest Model are then used to find the percentage of Weekdays and weekend when the load is light, medium or maximum.

Methodology:

Before using any of the techniques, it was important to clean the data. In the preprocessing step, mainly the following steps were followed:

1) **Null/ NAN values:** Any row having any null/ NAN value was discarded.
2) **One Hot Encoding:** Several columns including WeekStatus, Load_Type were categorial variables, which were converted into the respective encoded variables.
3) **Date Time:** The date time column was converted to date time object from the string datatype.

Data Visualization: This step was performed to get a general trend of the data, how the data was distributed. The density function and the scatter matrix were used for visualization. The scatter matrix shows the relation of various features with output, whereas density function shows the distribution of the data (which is mainly normal for almost all features).

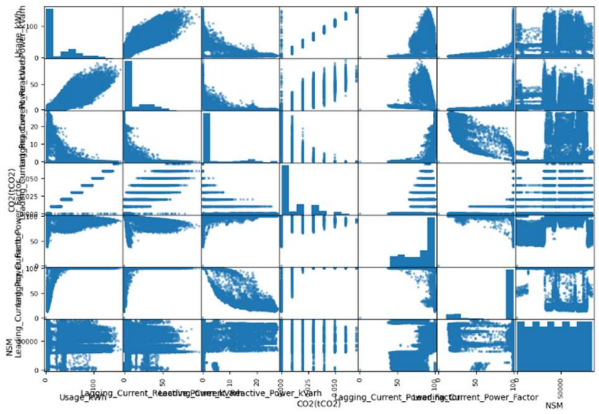

Fig. 2: Scatter Matrix

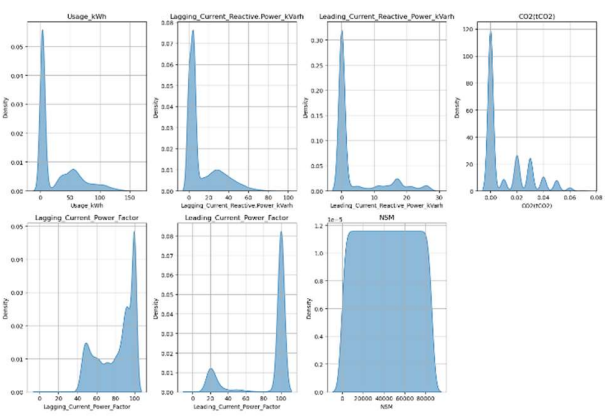

Fig. 3: Density Plot

ML Models:

1) **Linear Regression:** Linear Regression is generally used to predict rather than classification, however in this case, based on the value obtained, a certain threshold was defined for being to a particular category (0.5 in this case: if the predicted value >0.5, it belongs to the given class, else not). The dataset was divided into testing and training, which was used to further implement the model, using the following equation:

$$y = X\theta + \epsilon \quad (1)$$

The hyperparameter to be used was taken as the "Number of Variables Used". A general trend in decrease in RMSE was observed to take place as the number of variables increased. The confusion matrix suggested that approximately 15% of the samples were either False Positives or False Negatives.

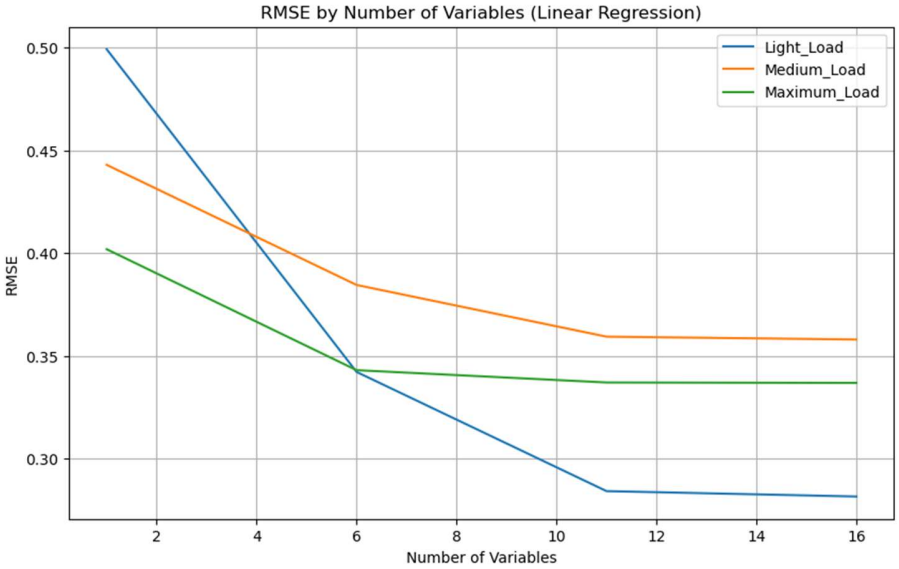

Fig. 4: Plot of RMSE against Hyperparameter (Linear Regression)

2) **SVM:** The use of SVM for prediction is comparatively more computational extensive, so the dataset was rather divided into a smaller fraction, consisting only 10% of the total dataset. Once the dataset is reduced, it is portioned into training and testing and the SVM models are constructed separately for the prediction of light, medium and maximum load types.

   In this case also, the number of parameters was taken as the hyperparameter, however in this case a different trend was observed, RMSE first decreased with an increase in the number of variables, however it then increased further.

   The SVM approach was also evaluated using the Confusion Matrix, however in this case also, ~10-15% of the data was wrongly predicted.

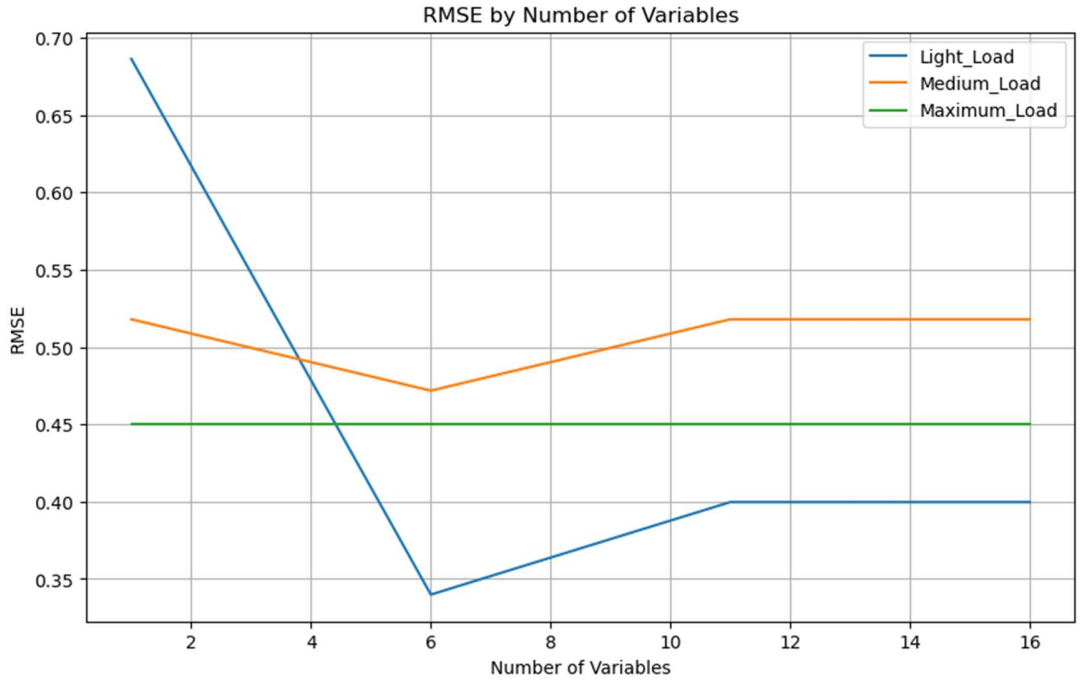

Fig. 5: Plot of RMSE against Hyperparameter (SVM)

3) **Random Forest [3]:** After Linear Regression and SVM, Random Forest Technique was used for prediction. In this technique also, the dataset was reduced in size, taking only 10% of the data. After partitioning of the dataset, the Random Forest Model was implemented.

   Again, the number of variables considered was taken as the hyperparameter. RMSE value decreased as the number of variables decreased. The final RMSE was much lesser than the other two models.

   Evaluation of this model using Confusion Matrix showed that the number of False Positives and False Negatives is comparatively lesser than the other 2 models, SVM and Linear Regression.

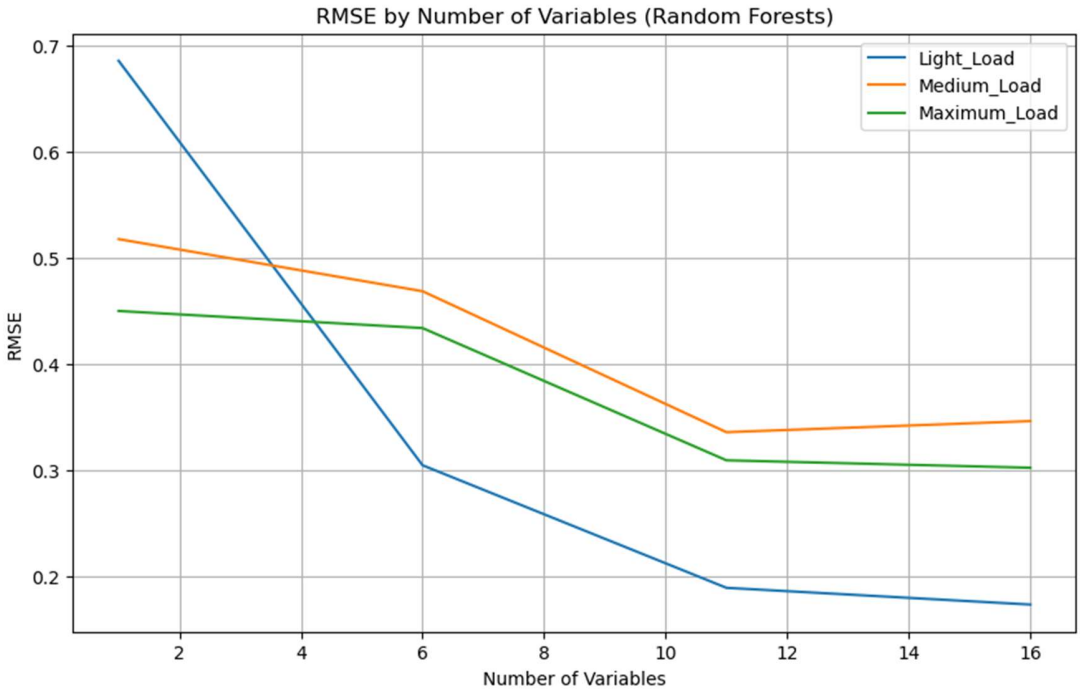

Fig. 6: Plot of RMSE against Hyperparameter (Random Forest)

Results:

The best model, i.e. the "Random Forest" was used to predict the trend of types of loads against the Week Status, which is one of the most important trends for prediction of load.

The last part of the code showed that when most of the offices weren't in use, i.e. on Weekends, the load was mostly light, whereas medium and maximum loads were observed on Weekdays.

```
Percentage of Light Load for Weekdays: 44.572158365261814
Percentage of Medium Load for Weekdays: 31.673052362707537
Percentage of Maximum Load for Weekdays: 23.754789272030653

Percentage of Light Load for Weekends: 69.15064102564102
Percentage of Medium Load for Weekends: 17.628205128205128
Percentage of Maximum Load for Weekends: 13.221153846153847
```

Fig. 7: Results using Random Forest Model

Conclusion:

This report focused on the description of the models predicting load types (Light, Medium, or Maximum Load) using machine learning models on data which was taken from the Korea Electric Power Corporation. Three models were used for this prediction: Linear Regression, Support Vector Machines (SVM), and Random Forests which were applied and evaluated based on their Root Mean Squared Error (RMSE) and confusion matrix results. Linear Regression exhibited a decreasing RMSE trend with more variables, with around 15% misclassification. SVM showed

mixed RMSE trends and misclassification rates of approximately 10-15%. Random Forests turned out to be the most effective model, which had lower RMSE as compared to other two models.
The "Random Forest" model, being the most effective, was further used to predict load trends about the WeekStatus, revealing that Weekends primarily experienced light loads, while Weekdays exhibited a mix of medium and maximum loads. This analysis provides valuable insights into efficient load prediction methods, essential for the sustainable development of smart cities, aimed at optimizing energy usage and enhancing living standards.
(To run the code, run all the blocks in the order they are present in the code)

References:

[1]: Seyedzadeh, S., Rahimian, F., Glesk, I. *et al.* Machine learning for estimation of building energy consumption and performance: a review. *Vis. in Eng.* **6**, 5 (2018)
[2]: Yongyun Cho *et al.* Efficient energy consumption prediction model for a data analytic-enabledindustry building in a smart city (2021)
[3]: Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001)
[4]: Amit Chawla. Predicting energy consumption through machine learning (2023)