A Comparison of Ensemble Methods for Predicting Power Consumption with Variation in Weather Parameters

Project Guide
Dr. Steffen Heber

Team Members
Abidaan Nagawkar (ajnagawk)
Abhishek Kedia (akkedia)
Albert Egido De Poy (aegidod)
Ankita Pise (aapise)
Shruti Gandhi (sgandhi3)
Vamshi Guduguntla (gudugu)

Project Description

Power consumption in a household is influenced by the weather conditions in a region. The ability to predict the consumption of power, based on variations in weather parameters, can enable the development of weather adaptive power factor correcting capacitor banks, which when installed in power distribution systems can significantly improve power efficiency. In this project, we train regression models on the power-weather training dataset, and perform predictive analysis on the test dataset using mean square error as the performance metric.

The technical objectives of this project are as follows:

- To build an ensemble method to predict the power consumption of the household with the variation in weather parameters.
- To compare the performance of the ensemble method against the constituent models and two other contemporary ensemble methods – AdaBoost and RandomForest.

Three different regression models were used as base inducers for the ensemble method:

- Linear Regression
- Regression Tree
- Support Vector Machine (SVM)

Project Motivation

The overall power factor of modern industries and power plant distribution system is very poor because of variable inductive loads absorbing reactive power with extreme variations in weather. These industries and power distribution systems require weather-adaptive automatic power factor correcting capacitor banks, improving the transmission capacity and the network stability. We aim to deploy this intelligence for smart capacitor banks for power factor correction. Another major motivation is to combine several weak models to produce a powerful ensemble.

Dataset Description

Data Sources:

- Individual household electric power consumption data set by UCI. [1]
- Weather Archive for Clamart, France available on the site. [2]

Data Description:

The two datasets obtained from the sources mentioned above were cleaned, aggregated and finally merged to get the final dataset to be used. The detailed step by step pre-processing of the datasets is explained in the Appendix. [3] This dataset contains the power consumption and weather readings on a daily basis for a period of three years. The dimensions of the dataset are 1023 observations of 11 variables each, including the power factor. The attributes characterizing each record are:

- Date_time, TemperatureF, DewpointF, PressureIn, WindDirectionDegrees, WindSpeedMPH, WindSpeedGustMPH, Humidity, HourlyPrecipIn, Dailyrainin, Power_Factor.

Flowchart

A flowchart detailing the steps carried out for the regression analysis using ensemble methods is given below:

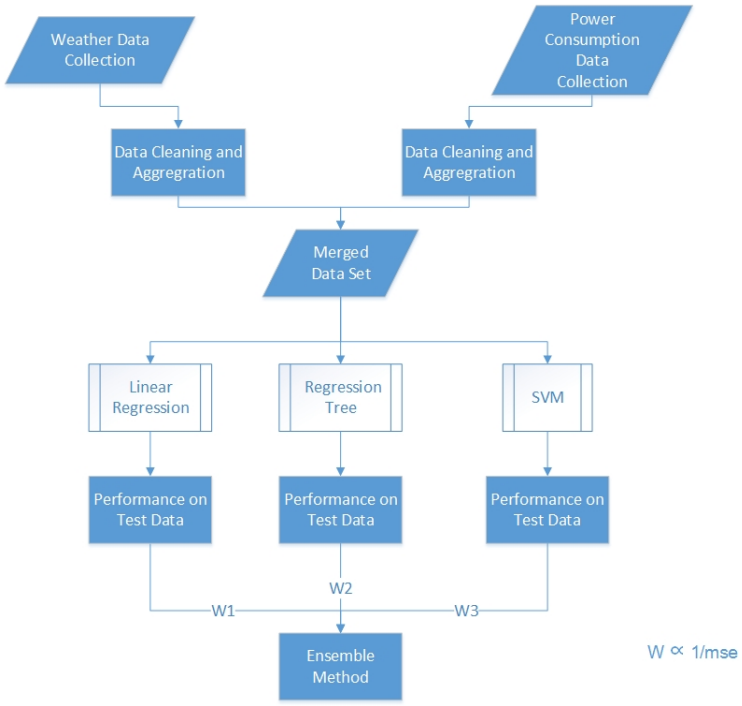

Figure 1: Flowchart for Predictive Ensemble Analysis

Exploratory Data Analysis:

An exploratory analysis was carried out to compare the contribution of each variable towards the regression model building. This was done using the RandomForest model and plotting the graph of variable importance. The results were as follows:

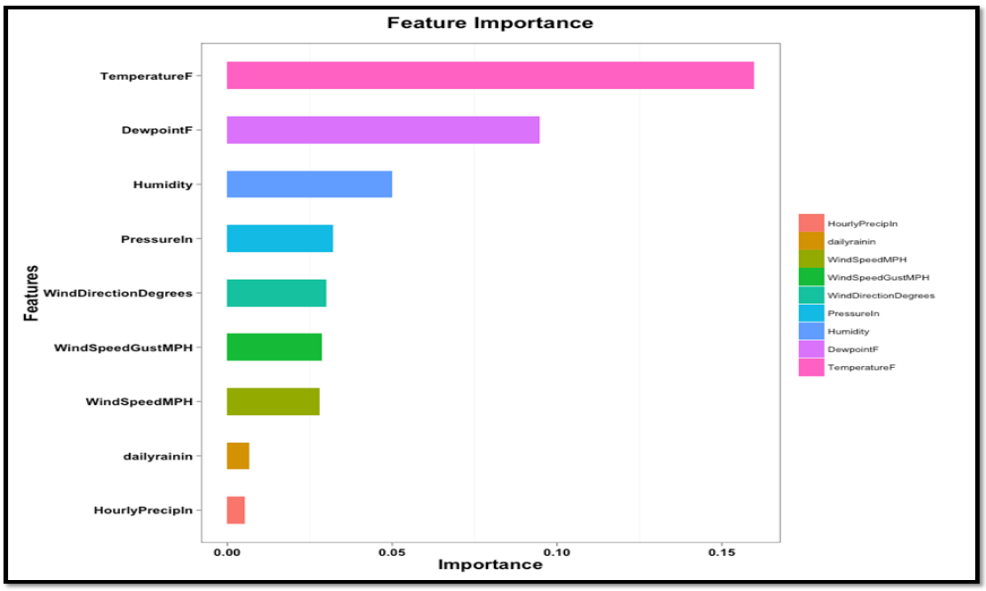

Figure 2: Feature Importance in the Weather Dataset

Implementation

After data was pre-processed and prepared for analysis, ensemble modeling was carried out in the following manner:

Constituent Regression Models: The regression models used as base inducers were Linear Regression, Support Vector Machine and Regression Tree model. The reasons for selecting each constituent regression model is as follows -

1. Linear Regression - The dataset was tested for the extent of association between the independent variables and the class label using correlation. High correlation between the weather parameters and power factor was the reason for selecting this method.
2. Regression Tree - The continuous target variable power factor, made this the most fundamental method of use for performing regression analysis.
3. SVM - SVM can be used to avoid difficulties of using linear functions in the high dimensional feature space and optimization problem is transformed into dual convex quadratic programs. In regression case the loss function is used to penalize errors that are greater than threshold - ε . Such loss functions usually lead to the sparse representation of the decision rule, giving significant algorithmic and representational advantages.[4]

Ensemble Method: The ensemble modeling was carried out using the 'caretEnsemble' package. The method uses the above three regression models as base inducers, and generates the final model by the linear greedy optimization of weights assigned to each method depending on their respective mean square error for 25 iterations. The performance for all the methods is as shown below -

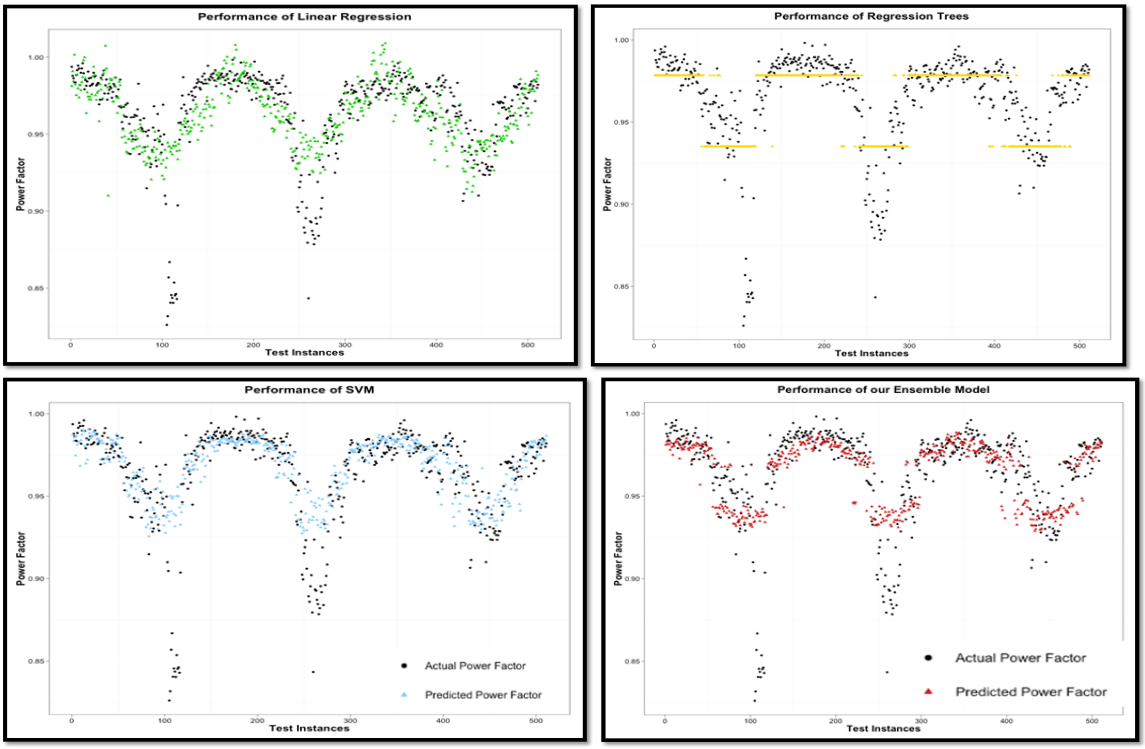

Figure 3: In clockwise order; Performance of Linear Regression, Performance of Regression Tree, Performance of our Ensemble Method, Performance of SVM

Comparison with AdaBoost and RandomForest: Here we wanted to compare the performance of our ensemble method with both random forest which performs well typically for deep trees (level >=7), as well as AdaBoost, which typically works better with shallow trees (5-15 leaves). [6]

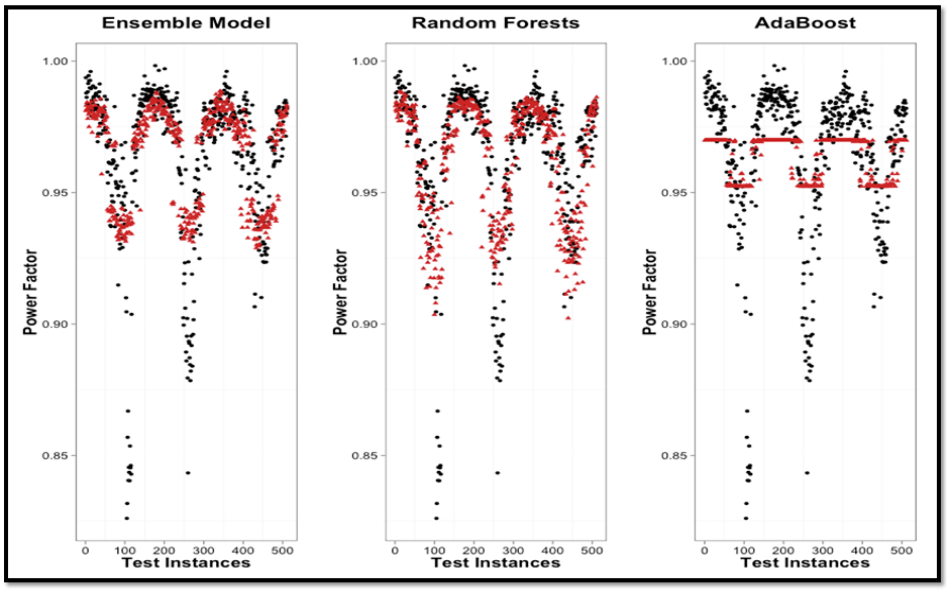

Figure 4: Comparison of our Ensemble method with Random Forest and AdaBoost

Results

Performance metric: Mean Square Error

The graph below compares mean square error for all the regression methods used in this project. Highlighted in red is the variance of all the methods from the Ensemble method. Our proposed method shows maximum improvement over the Adaboost method, and is comparable in performance with the Random Forest.

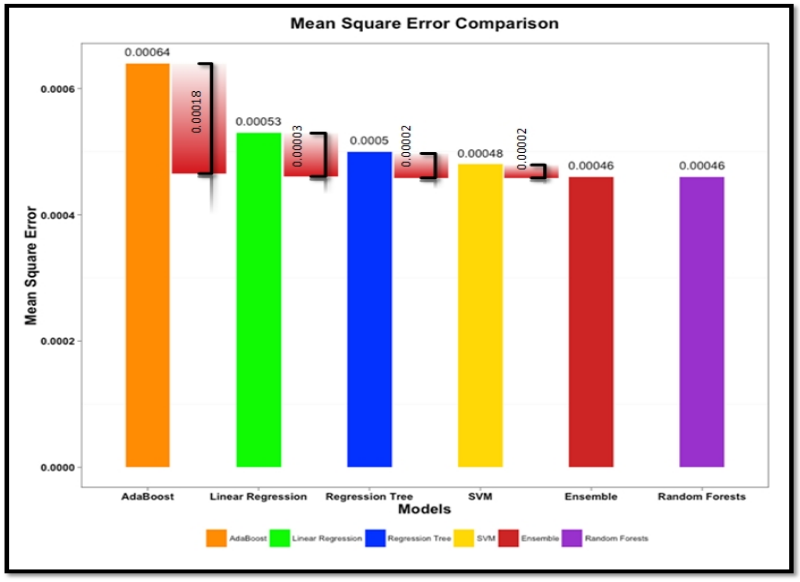

Figure 5: Mean Square Error Comparison for AdaBoost, Linear Regression, Regression Tree, SVM, Ensemble Method, Random Forest (Left to Right)

Discussion

Challenges:

1. One of the main challenges we faced was the creation of the dataset. We had access to the power dataset, but had to write a script to download the weather dataset.
2. Another interesting challenge was the aggregation of these datasets and their merging. The power dataset had a granularity of one minute while the weather dataset had a granularity of five minutes. We aggregated these datasets to two granularities: a) One hour and b) One day. The merging of the two datasets then became a trivial task.
3. The third major challenge we faced was building the actual ensemble model. Initially, we tried assigning weights to the individual models ourselves, but this did not always result in the best model. We then had to use the caret and caretEnsemble libraries in order to build both, the individual models as well as the ensemble model.
4. Another challenge was, understanding how to use the ggplot2 library. One change that we would like to make, given the chance, is to build our own theme for the various plots, instead of setting those parameters manually for each plot.

In terms of **achievements**, we achieved a mean squared error comparable to the one for Random Forests, which is considered to be one of the best techniques currently in use.

Future Scope

As future work, it would be interesting to perform some Time Series Analysis and compare the results with the ones obtained with our ensemble method. Since the data points are highly correlated we could fit an Autoregressive (AR) model and evaluate the output with future data. Our method can also be extended for carrying out predictive analyses for many households. This can also be a possible area to be explored in the future.

References

[1] Individual household electric power consumption data set by UCI - https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption
[2] Weather History Data for LFPO, nearest airport to Clamart, France - http://www.wunderground.com/history/airport/LFPO/2015/2/26/DailyHistory.html?req_city=Clamart&req_statename=France&reqdb.zip=00000&reqdb.magic=15&reqdb.wmo=07156
[3] Appendix A
[4] http://kernelsvm.tripod.com/
[5] Bauer, E., Kohavi, R., An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants (1999)
[6] Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference and Prediction", Second Edition, Chapter 15 (page 589 - 591)