Multivariate-analysis of Predictors influencing Electricity and Water used by Commercial Buildings in the U.S.

Data Set: Commercial Buildings Energy Consumption Survey(CBECS) – 2012 by Energy Information Administration(EIA); Commercial Buildings Water Usage Survey(CBECS) – 2012 by Energy Information Administration(EIA)

Data Preparation and Variable Imputation

Energy usage Data: 6720 observations of 1119 variables

Water usage Data: 1129 observations of 214 variables

On exploring both the datasets we observed that water usage survey was carried out only for large commercial buildings greater than 200,000 square feet. In order to prepare the data for the multivariate study, we did a inner join between two datasets on variable ‘PUBID’. The merged dataset had 1129 observations with 1121 variables

The original CBECS 2012 data had a lot of values that needed to be imputed. Of the null values in the data we imputed the data based on the imputation flags provided by EIA for the specific variables.

We also made some specific changes to certain variables to make them more “R-friendly”. Those details are listed below:

1) Based on initial hypothesis formed, 158 variables of interests were considered from the merged dataset.

2) All null values were cross checked with the imputation flags for the same variable. The imputation flags indicated that there were two types of values which were marked NA. These were either “Data not Provided” or “Not Applicable to that Building”. The NA values were filled in with these as factor variables.

3) Variables with greater than 40% missing values were dropped. A lot of missing values refer to the incomplete data filled during the survey. This reduces variables in the model to 92 variables

2) Percentages in many places in this data set were used in conjunction with another variable having Yes/No data. This information was used to impute missing values in some cases

3) NFLOOR was used to indicate the number of floors of the building. This was further transformed as factor since it contained values as 994,995

4) Variable imputation for missing numerical variables was done using ‘mice’ package and imputed variables showing good resemblance with original distribution were replaced with the missing values.

5) Observations with missing values for important predictors (‘BLDSHP’,’NELVTR’) were removed as there were not good imputations values for the variables

Data Description and Major Predictors

Annual electricity consumption (kWh)– ELCNS
Annual water consumption (gallons)- WTCNS

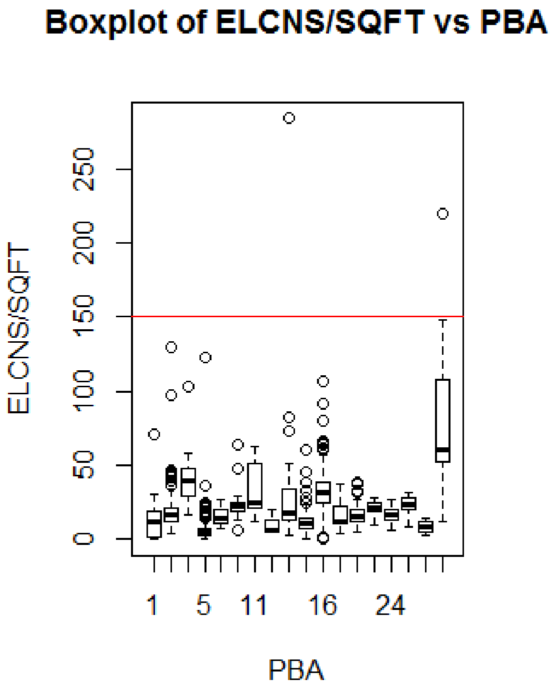
Important Predictors are:

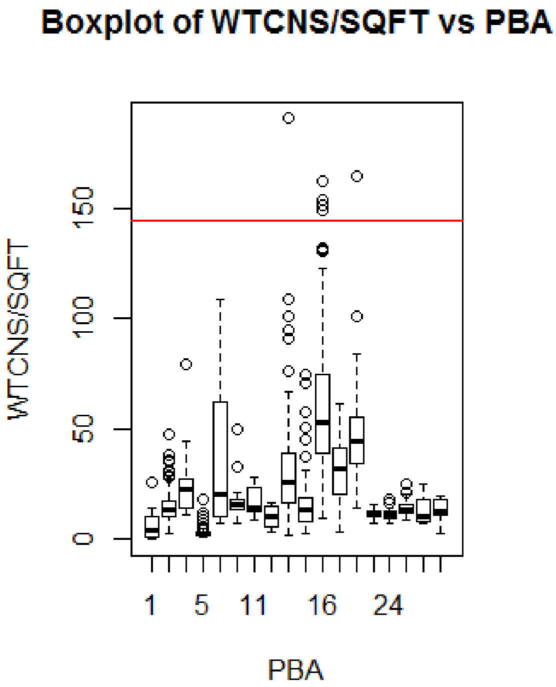
- Square footage-SQFT
- Principal building activity-PBA
- Main heating equipment-MAINHT
- Census Division-CENDIV
- Number of employees -NWKER
- Number of Elevators- NELVTR
- Number of floors-NFLOOR
- Year of construction category-YRCONC
- Number of computers -PCTERMN
- Cooling Degree Days-CDD65
- Number of laptops -LAPTPN
- Building shape-BLDSHP
- Main cooling equipment-MAINCL
- Lit off Category- LNHRPC
- Percent exterior glass-GLSSPC
- Major fuel heating use (thous Btu) MFHTBTU
- Total hours open per week-WKHRS

Variable selection, Outlier Treatment

Initial variable selection was carried graphically on the basis of violin plots and scatter plot for categorical and numerical variables. Post that we also fitted a multivarite tree boosting models to estimate importance variables. This helped us reduce to 32 variables from 92 variables initially taken(Graphs are provided in the appendix section)

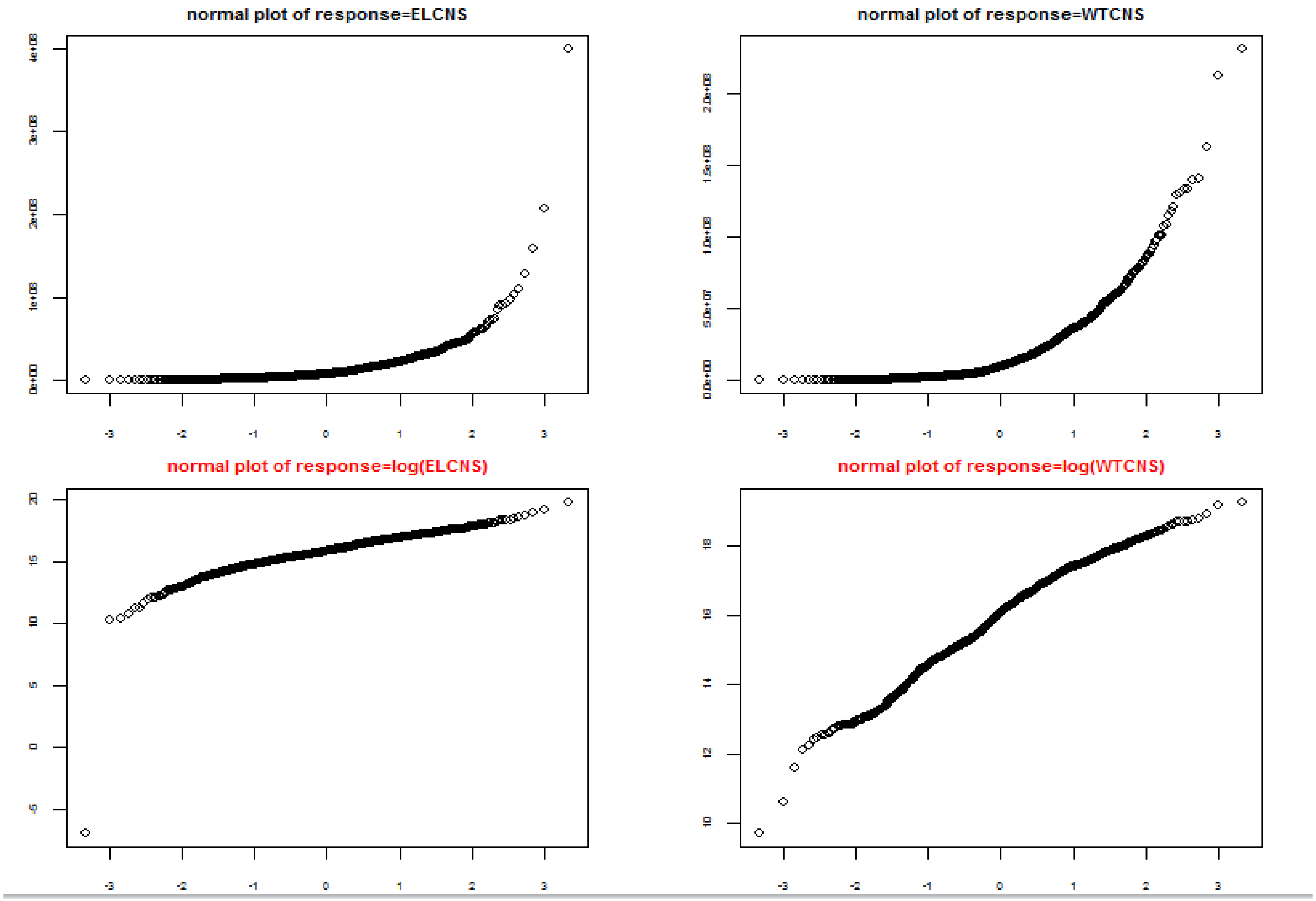
There were some observations with extreme energy(Electricity/Water) Intensity= ELHTBTU/SQFT. We removed the observations with these extreme values.





Data Visualization

We aim to build good predictive models by looking at key characteristics (Normality, Variance, correlation) of our predictors and response variable

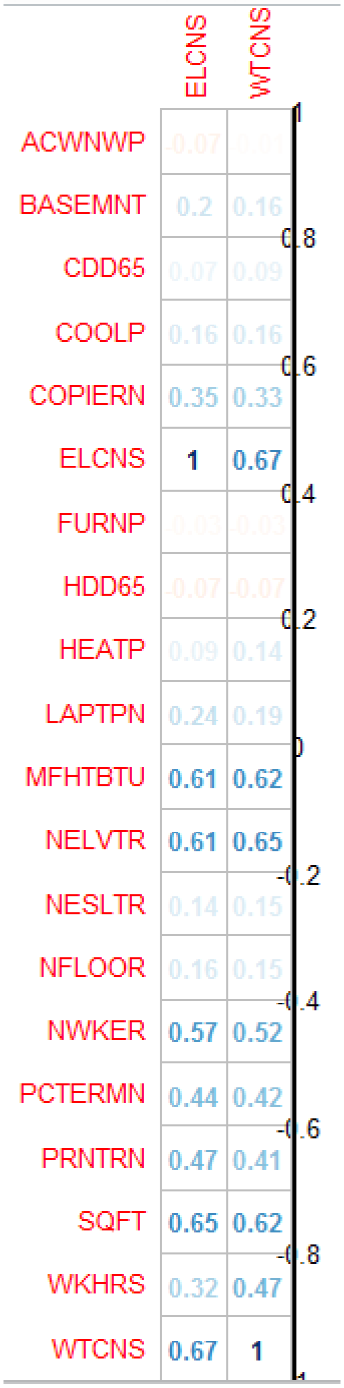


The QQ-Plot of the original response variable shows non-normal behavior. The log plots clearly indicate that transferring the response variable-WTCNS, ELCNS to the log scale significantly improves the frequency distribution, also the QQ plot indicates normality.

Log transformation is a good idea for the response variable, considering the original values are very widespread.

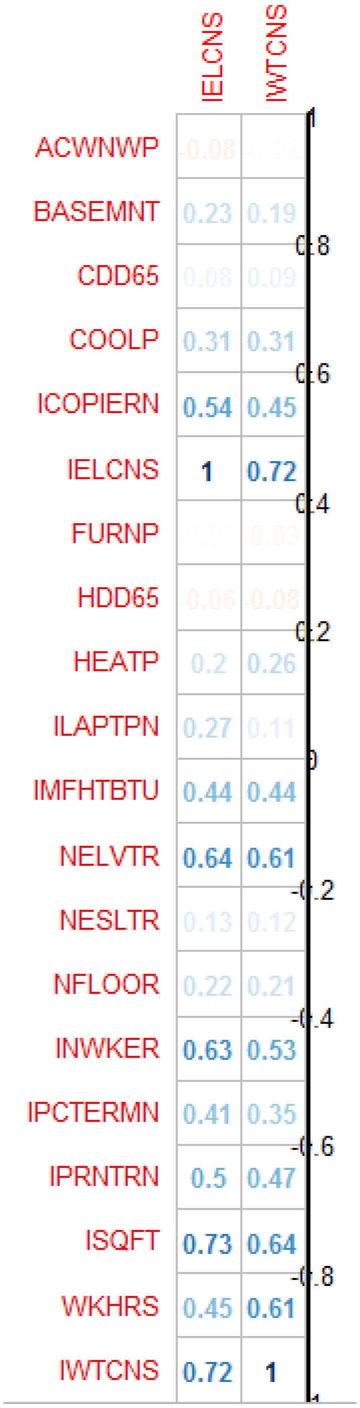
Based on the same logic we have transformed 'MFHTBTU','DHBTU','ELCNS','COPIERN','PRNTRN','PCTERMN','WTCNS','LAPTPN','NWKER' since it improves correlation with lELCNS,lWTNS

Log Transformations and Correlation analysis



	ELCNS	WTCNS
ACWNWP	-0.07	-0.01
BASEMNT	0.2	0.16
CDD65	0.07	0.09
COOLP	0.16	0.16
COPIERN	0.35	0.33
ELCNS	1	0.67
FURNP	-0.03	-0.03
HDD65	-0.07	-0.07
HEATP	0.09	0.14
LAPTPN	0.24	0.19
MFHTBTU	0.61	0.62
NELVTR	0.61	0.65
NESLTR	0.14	0.15
NFLOOR	0.16	0.15
NWKER	0.57	0.52
PCTERMN	0.44	0.42
PRNTRN	0.47	0.41
SQFT	0.65	0.62
WKHRS	0.32	0.47
WTCNS	0.67	1

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



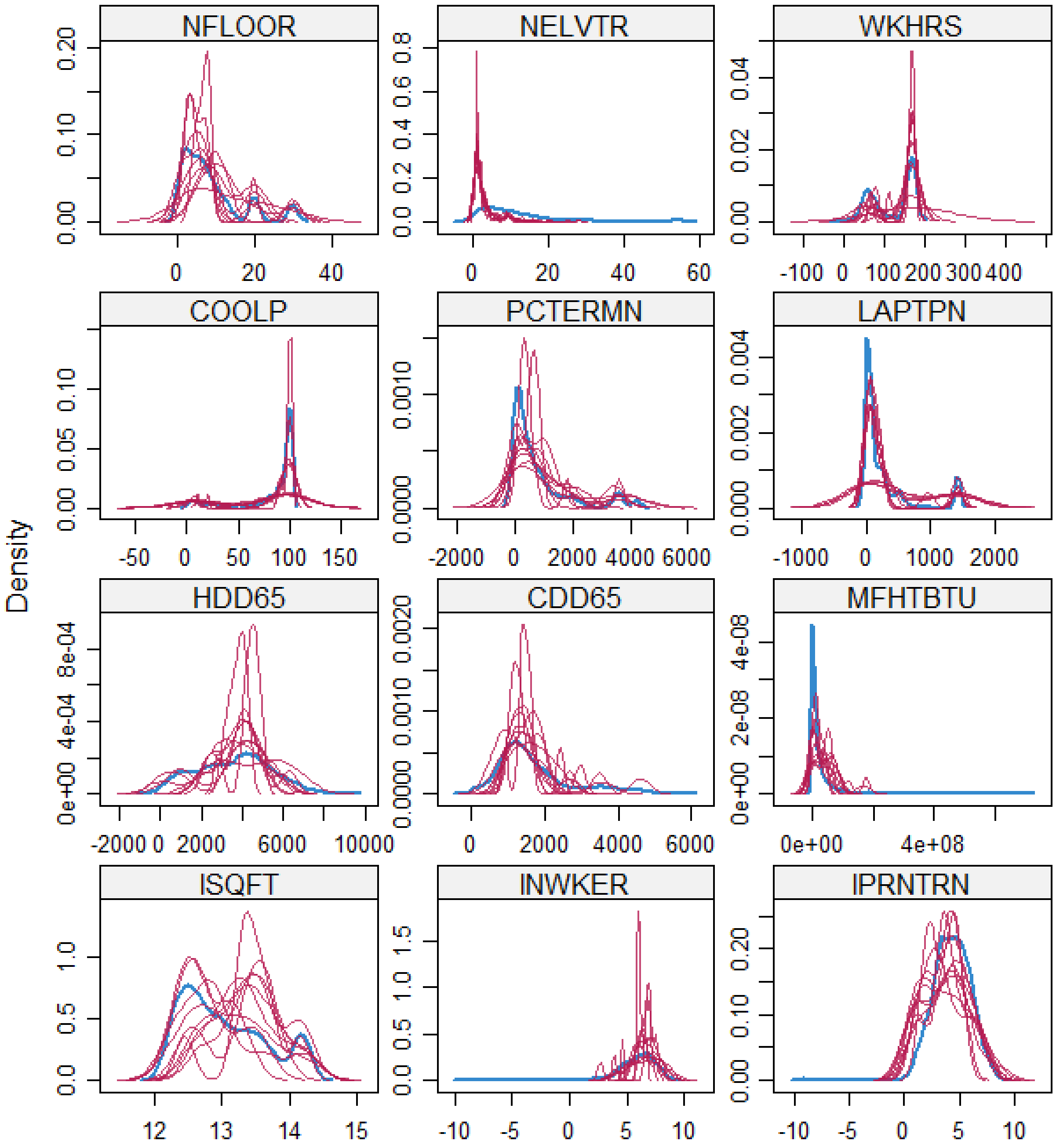
	IELCNS	IWTCNS
ACWNWP	-0.08	-0.02
BASEMNT	0.23	0.19
CDD65	0.08	0.09
COOLP	0.31	0.31
ICOPIERN	0.54	0.45
IELCNS	1	0.72
FURNP	-0.01	-0.03
HDD65	-0.06	-0.08
HEATP	0.2	0.26
ILAPTPN	0.27	0.11
IMFHTBTU	0.44	0.44
NELVTR	0.64	0.61
NESLTR	0.13	0.12
NFLOOR	0.22	0.21
INWKER	0.63	0.53
IPCTERMN	0.41	0.35
IPRNTRN	0.5	0.47
ISQFT	0.73	0.64
WKHRS	0.45	0.61
IWTCNS	0.72	1

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Based on above output log transformations of response variables shows improved correlations and normal distribution. PCTERMN, LAPTPN, MFHTBTU were later removed from log transformations as these variables didn’t show improvement in correlations and normal behavior

Variable imputation

Missing values in numeric variables were imputed using ‘mice’ package. The method chosen was ’rf’ and iterations =10



Variable NELVTR are having 147 missing values in this case. Rest Numerical variables have less than 1% missing values(8-10 in the dataset). Therefore except NELVTR(imputed distribution didn’t resemble well the original distribution); numerical variables are imputed using above dataset since imputed Dataset able to model original distribution in each case .

Methodology

Boosted decision tree ensembles (Friedman, 2001) are a powerful off-the-shelf learning algorithm, allowing dependent variables to be non-linear functions of predictors as well as handling predictors with missing data. In addition to having high prediction performance, Boosted decision trees are an extremely flexible approach for exploratory data analysis.

One of the challenges of using multivariate decision tree ensembles is that the model is more difficult to interpret than a single tree. While tree boosting can be used to build a very accurate predictive model, it is potentially more important for researchers to interpret the effects of predictors. Below, we describe approaches that have been developed to

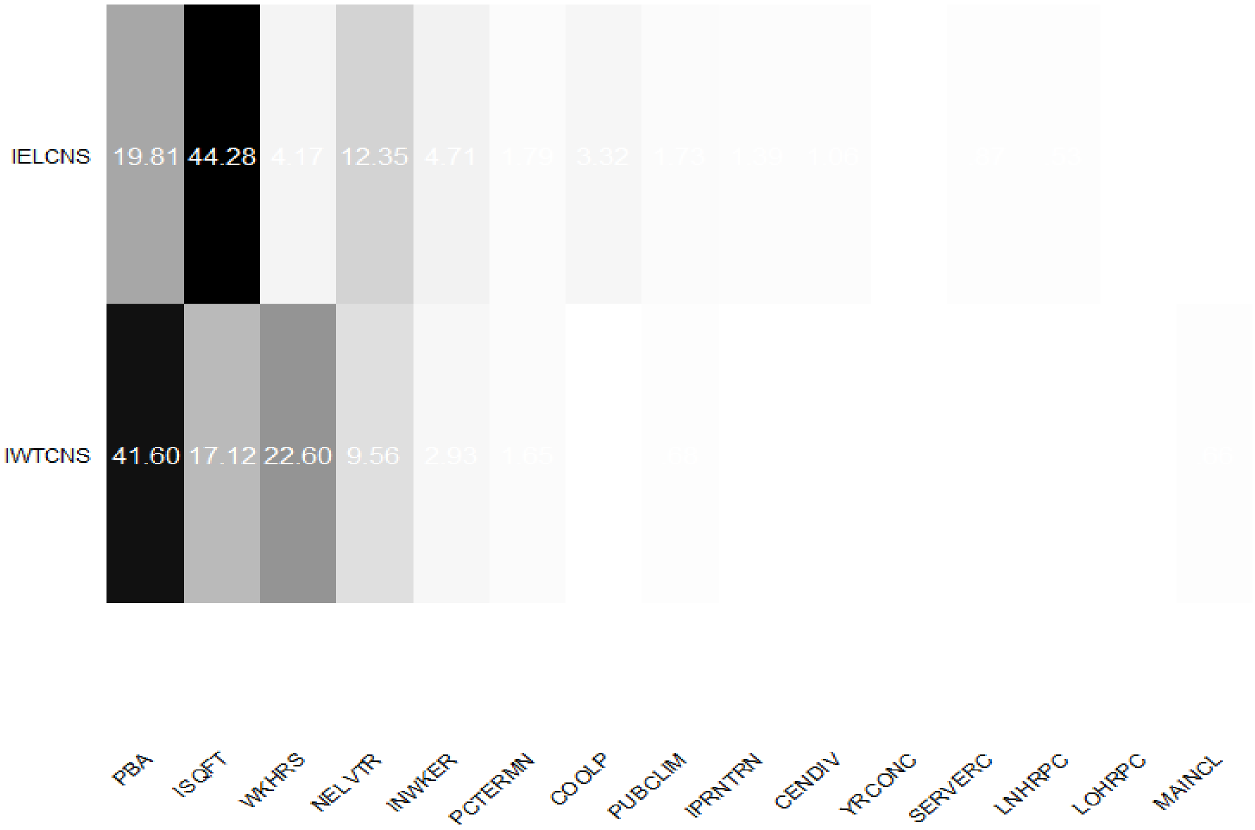
- identify predictors with effects on individual outcome variables
- identify groups of predictors that jointly influence one or more outcome variables
- visualize the functional form of the effect of important predictors
- detect predictors with possible interaction non-linear effects.

A well known R package for fitting boosted decision trees is gbm. This package extends gbm to multivariate, continuous outcome variables by fitting a separate univariate model of a common set of predictors to each outcome variable. This common basis accounts for covariance in the outcome variables as in seemingly unrelated regression. We refer to the package gbm and the extensive literature on boosting with decision trees for theoretical and technical details about how such a model is fit and interpreted (see references below).

Relative influence

In gradient boosting, the number of trees, the shrinkage, and the tree depth are meta-parameters that are important to tune to improve the fit of the model. Typically, the shrinkage is fixed to a small value, and the optimal number of trees is chosen by cross-validation.

The influence (or variable importance) of each predictor can be used to identify ‘important’ predictors. It is defined as the reduction in sums of squared error due to any split on that predictor, summed over all trees in the model (Friedman, 2001). Usually the score is relative, expressed as a percent of the total sums of squared error reductions from all predictors.



The output shows that ‘PBA’,’lSQFT’,’WKHRS’,’NELVTR’,’lNWKER’,’PCTERMN’ ,’COOLP’,’PUBCLIM’ are important predictors that are impacting the response variables. We will further explore this with covariance explained by predictors

Fit

As a check of the overall fit of the model, the R2R2 in the test set can be computed for each dependent variable.

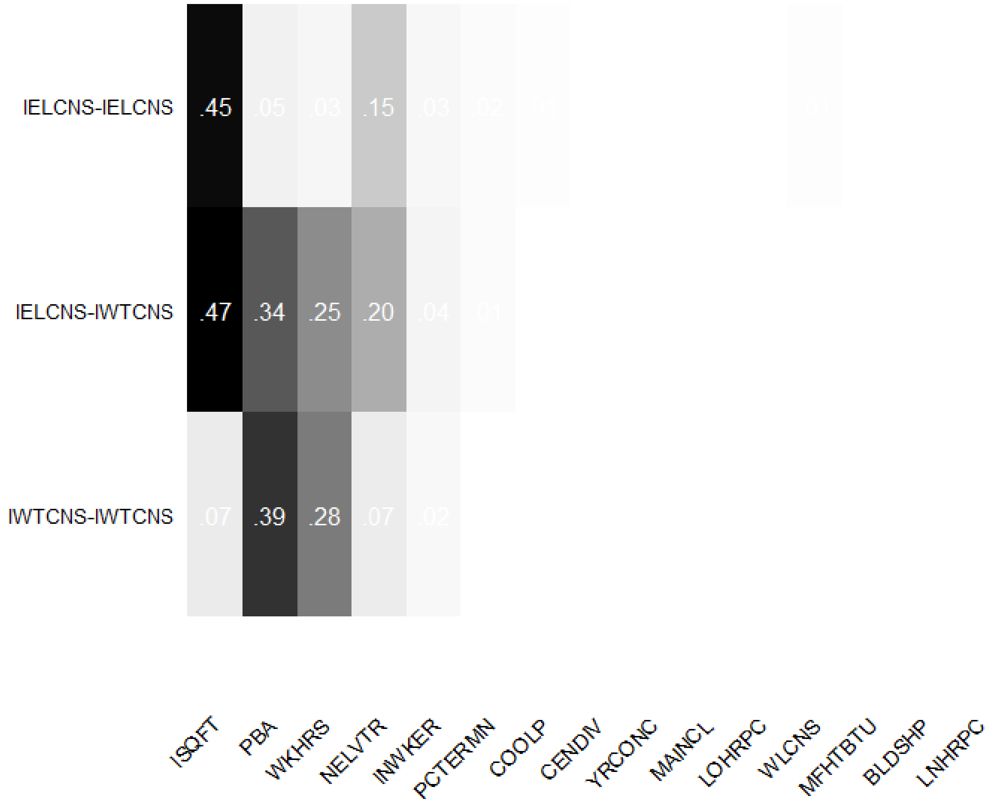
```
   lWTCNS    lELCNS
0.7498971 0.6094590
```

This shows that model is a good fit for Water usage; however it doesn’t explain much variation in Electricity usage.

Covariance explained

It may also be informative to select a set of outcome variables that are associated with groups of predictors. One criterion for selecting outcome variables is to choose the outcome variables whose covariance can be explained by a function of a common set of predictors. The covariance explained in each pair of outcomes by predictors is estimated directly in mvtb.

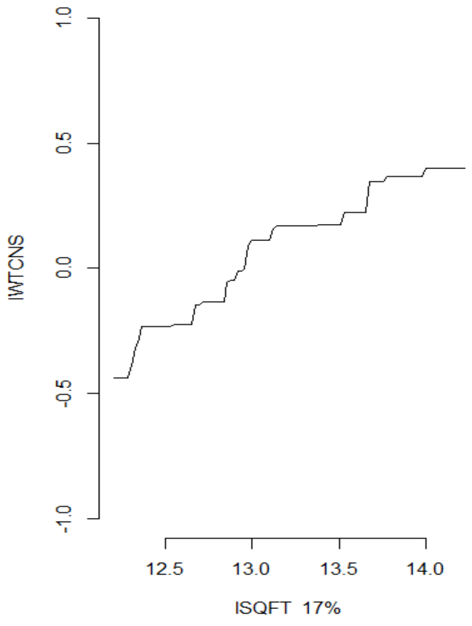
A covariance-explained matrix be organized as a Q(Q+1)/2×pQ(Q+1)/2×p table, where QQ is the number of outcomes, and pp is the number of predictors. Each element is the covariance explained in any pair of outcomes by a predictor. When the outcomes are standardized to unit variance, each element can be interpreted as the correlation explained in any pair of outcomes by a predictor.

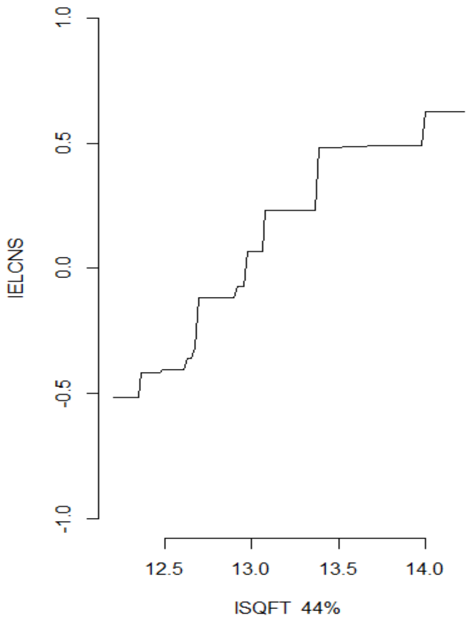


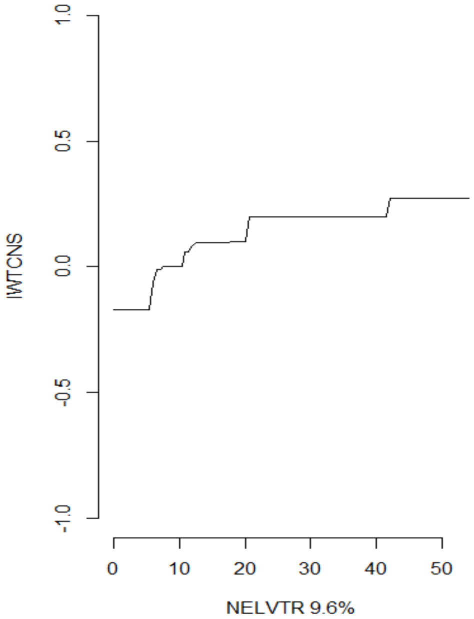
The predictors have similar influence as observed in relative influence section. Particularly, lSQFT,PBA explain majority of variation in both repsonses

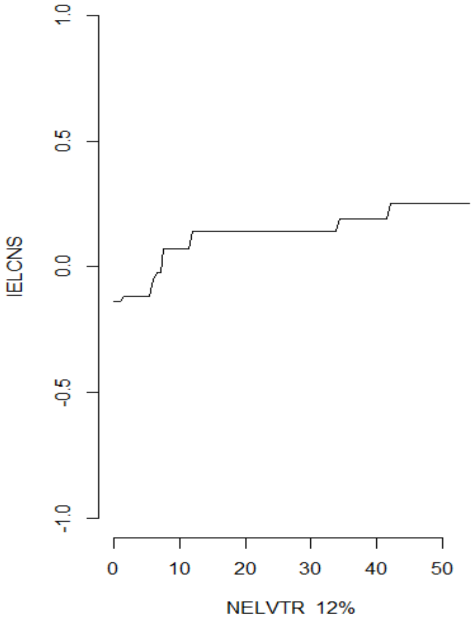
4. Partial Dependence Plots

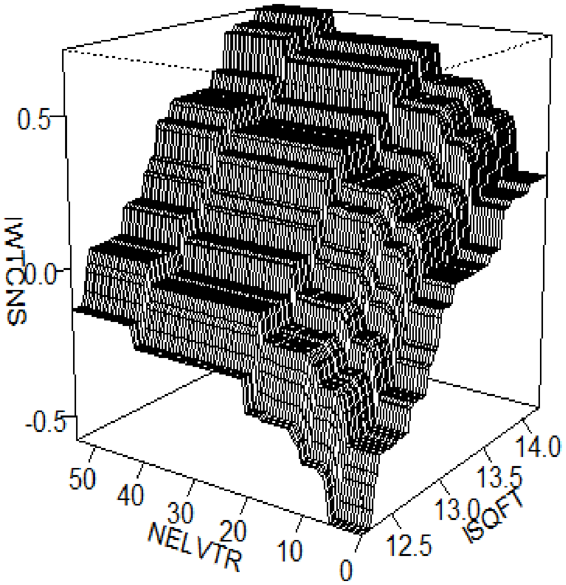
Partial dependence plots complement interpretations of relative influence by showing the direction and functional form of the effect of the predictor. A partial dependence plot shows the effect of the predictor averaging over (or integrating out) the effects of other predictors.

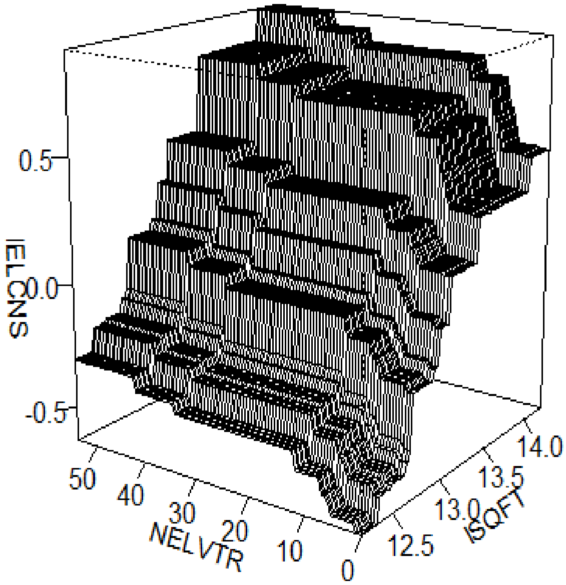












From above plots, we observe that Electricity and Water usage increases with increase in SQFT area. Similar pattern is observed with Number of elevators in the building .Both the observations are in line with initial hypothesis during data exploration.

The third plots shows how ‘lSQFT’ and ‘NELVTR’ interact together. For some areas in graph of ‘lELCNS’ vs ‘NELVTR’,’lSQDT’; we observe a non additieve relationships between the variables. This can be further tested numerically. Similar plots for important predictors are in appendix

Inferences

- Square Footage Area and Principal Building activity are the two major factors determining Electricity and water consumption in commercial buildings. Electricity and Water consumption tends to increase with increase in square footage area of building reflecting increase in number of resources used(equipment, staff etc).
- Buildings with purposes as 'Public Assembly, 'Inpatient Health Care' and 'Enclosed Wall tends to have a higher water and energy consumption that other building activities. Vacant buildings have a negative impact on energy and water consumption which reflects the practical situation. Also business places like offices and laboratory have a low consumption pattern compared to other purposes.
- Consumption pattern in buildings with active working hours less than 168 have lower consumption profile where as building with 168 working hours a week operations, have higher energy consumption pattern
- Number of Elevators and Number of Workers directly impacts energy and water consumption in commercial buildings
- Census divisions 6 and 7 (Southern divisions) show higher electricity and water consumption levels, which is pretty intuitive.
- Electricity consumption and water usage increases proportionally with increase in number of PCs and Cooling percentage

Departures from additivity

Below, we compute the top 5 departures for each Dependant variables.

```
$lWTCNS
  var1.index var1.names var2.index var2.names nonlin.size
1         29     lNWKER         12      WKHRS    9.155954
2         12      WKHRS         10     NELVTR    9.125244
3         29     lNWKER         10     NELVTR    7.777051
4         29     lNWKER         28      lSQFT    7.624989
5         28      lSQFT          2        PBA    7.575741
6         28      lSQFT         12      WKHRS    6.943622

$lELCNS
  var1.index var1.names var2.index var2.names nonlin.size
1         28      lSQFT          2        PBA    14.14408
2         29     lNWKER         28      lSQFT    13.95768
3         28      lSQFT         10     NELVTR    12.80581
4         28      lSQFT         12      WKHRS    12.16519
5         30    lPRNTRN         28      lSQFT    11.52197
6         28      lSQFT         14      COOLP    10.69552
```

Here we observe that for ELCNS, there is some departure from additivity for 'lSQFT' and 'NELVTR'. We also observed this in the previous graphs. Similarly analysis can be done for other variables too.

Appendix

Data: https://www.eia.gov/consumption/commercial/

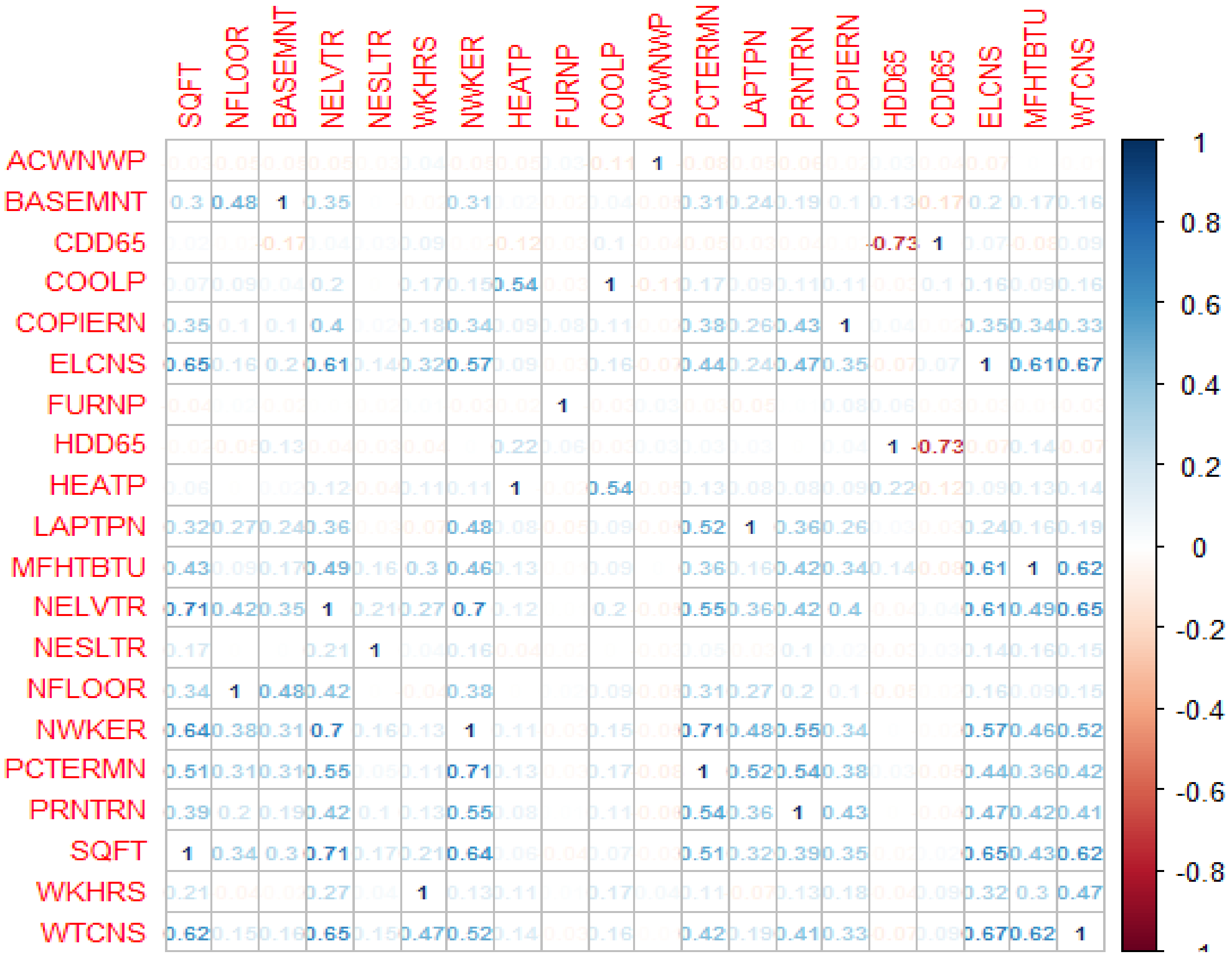
Log transformations: All numerical variables with their log transformation plots are saved in the pdf

some.pdf

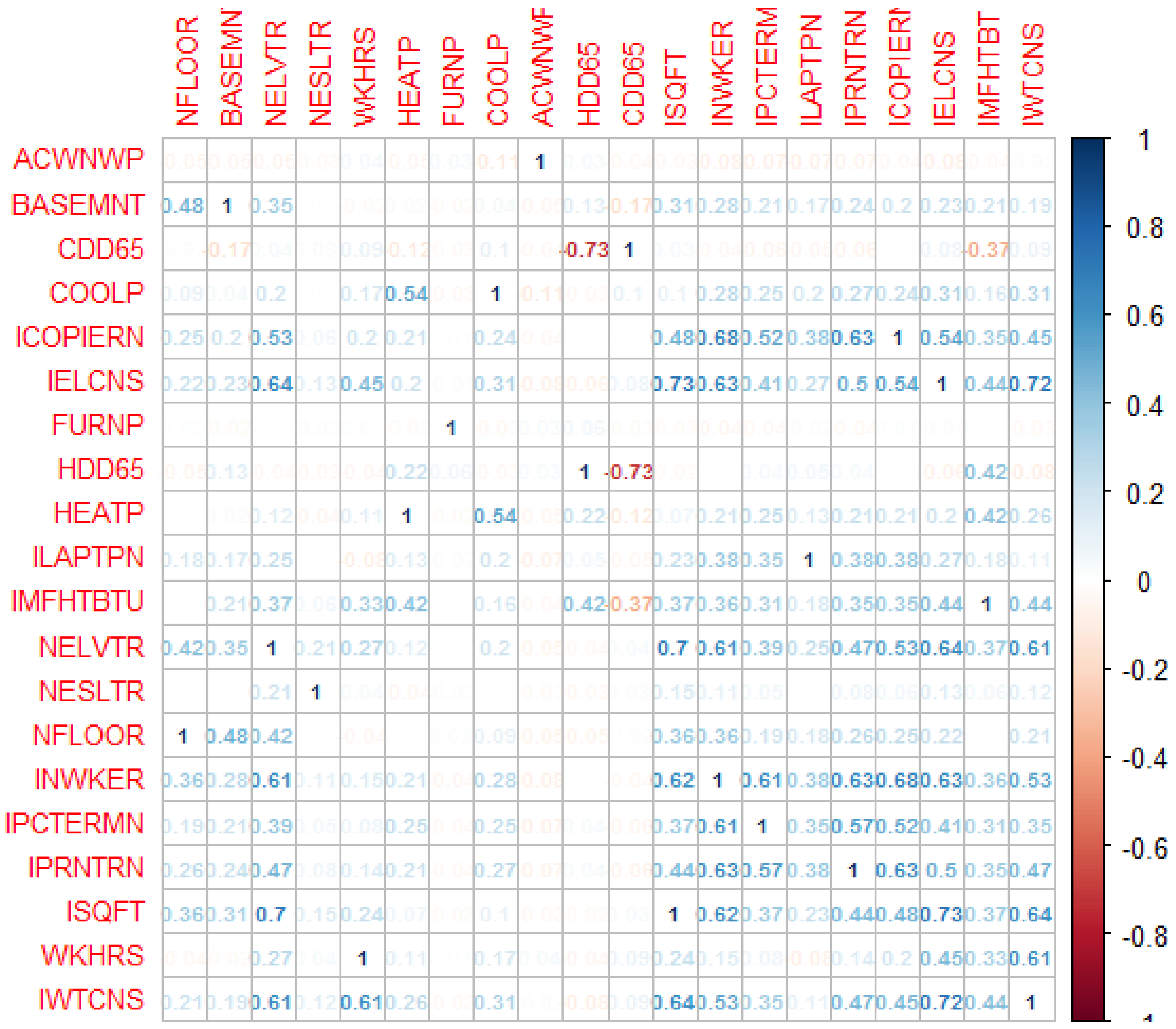
Based on these predictors were chosen for log transformations

Correlations

Correlation between untransformed variables



Correlation between transformed variables



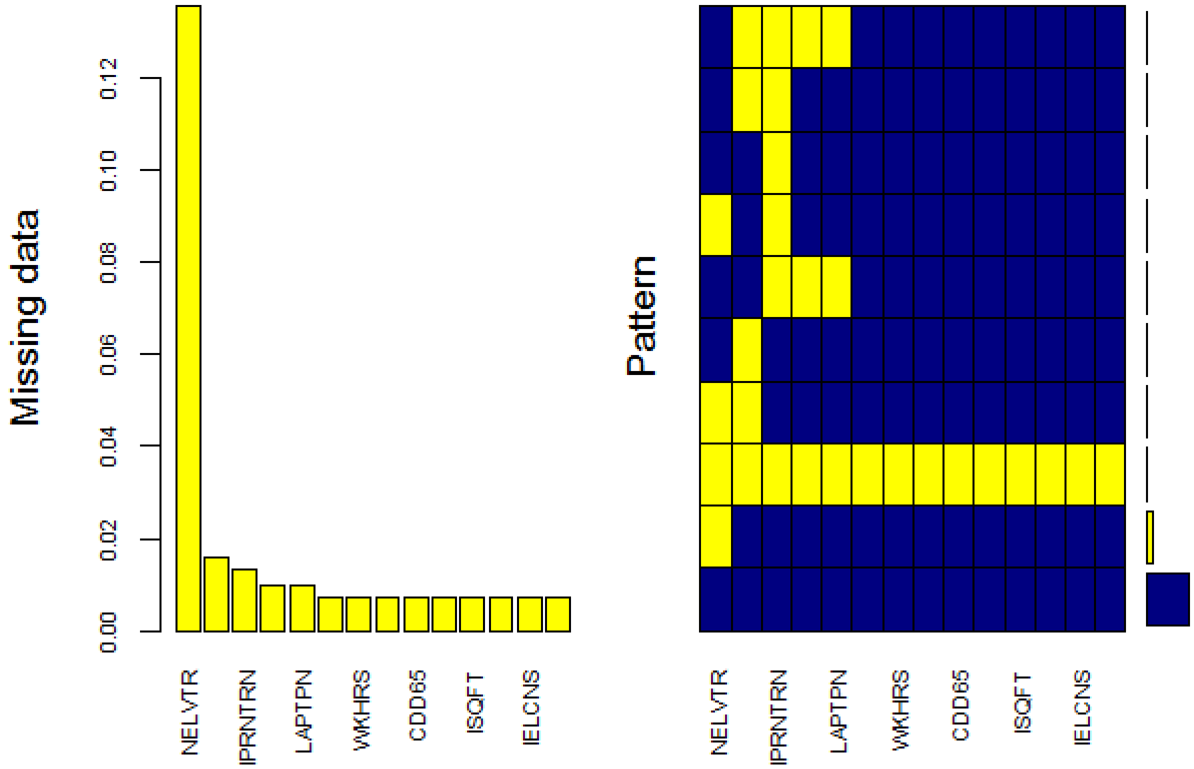
	NFLOOR	BASEMN'	NELVTR	NESLTR	WKHRS	HEATP	FURNP	COOLP	ACWNWF	HDD65	CDD65	ISQFT	INWKER	IPCTERM	ILAPTPN	IPRNTRN	ICOPIER	IELCNS	IMFHTBT	IWTCNS
ACWNWP	-0.05	-0.05	-0.05	-0.03	0.04	-0.05	0.03	-0.11	1	0.03	-0.04	-0.03	-0.08	-0.07	-0.07	-0.07	-0.04	-0.08	-0.04	-0.02
BASEMNT	0.48	1	0.35		-0.02	0.02	-0.02	0.04	-0.05	0.13	-0.17	0.31	0.28	0.21	0.17	0.24	0.2	0.23	0.21	0.19
CDD65	-0.02	-0.17	0.04	0.03	0.09	-0.12	0.03	0.1	-0.04	-0.73	1	0.03	-0.04	-0.06	-0.05	-0.06		0.08	-0.37	0.09
COOLP	0.09	0.04	0.2		0.17	0.54	-0.03	1	-0.11	-0.03	0.1	0.1	0.28	0.25	0.2	0.27	0.24	0.31	0.16	0.31
ICOPIERN	0.25	0.2	0.53	0.06	0.2	0.21	0.01	0.24	-0.04			0.48	0.68	0.52	0.38	0.63	1	0.54	0.35	0.45
IELCNS	0.22	0.23	0.64	0.13	0.45	0.2	0.01	0.31	-0.08	-0.06	0.08	0.73	0.63	0.41	0.27	0.5	0.54	1	0.44	0.72
FURNP	0.02	-0.02		-0.02	0.01	-0.02	1	-0.03	-0.03	0.06	-0.03	-0.03	-0.04	-0.04	-0.03	-0.04	0.01	0.01		-0.03
HDD65	-0.05	0.13	-0.04	-0.03	-0.04	0.22	0.06	-0.03	0.03	1	-0.73	-0.02		0.04	0.05	0.04		-0.06	0.42	-0.08
HEATP		0.02	0.12	-0.04	0.11	1	-0.02	0.54	-0.05	0.22	-0.12	0.07	0.21	0.25	0.13	0.21	0.21	0.2	0.42	0.26
ILAPTPN	0.18	0.17	0.25		-0.08	0.13	0.02	0.2	-0.07	0.05	-0.05	0.23	0.38	0.35	1	0.38	0.38	0.27	0.18	0.11
IMFHTBTU		0.21	0.37	0.06	0.33	0.42		0.16	-0.04	0.42	-0.37	0.37	0.36	0.31	0.18	0.35	0.35	0.44	1	0.44
NELVTR	0.42	0.35	1	0.21	0.27	0.12		0.2	-0.05	-0.04	0.04	0.7	0.61	0.39	0.25	0.47	0.53	0.64	0.37	0.61
NESLTR			0.21	1	0.04	-0.04	0.02		-0.03	-0.03	0.03	0.15	0.11	0.05		0.08	0.06	0.13	0.06	0.12
NFLOOR	1	0.48	0.42		-0.04		0.02	0.09	-0.05	-0.05	-0.02	0.36	0.36	0.19	0.18	0.26	0.25	0.22		0.21
INWKER	0.36	0.28	0.61	0.11	0.15	0.21	-0.04	0.28	-0.08		-0.04	0.62	1	0.61	0.38	0.63	0.68	0.63	0.36	0.53
IPCTERMN	0.19	0.21	0.39	0.05	0.08	0.25	-0.04	0.25	-0.07	0.04	-0.06	0.37	0.61	1	0.35	0.57	0.52	0.41	0.31	0.35
IPRNTRN	0.26	0.24	0.47	0.08	0.14	0.21	-0.04	0.27	-0.07	0.04	-0.06	0.44	0.63	0.57	0.38	1	0.63	0.5	0.35	0.47
ISQFT	0.36	0.31	0.7	0.15	0.24	0.07	-0.03	0.1	-0.03	-0.02	0.03	1	0.62	0.37	0.23	0.44	0.48	0.73	0.37	0.64
WKHRS	-0.04	-0.02	0.27	0.04	1	0.11	0.01	0.17	0.04	-0.04	0.09	0.24	0.15	0.08	-0.08	0.14	0.2	0.45	0.33	0.61
IWTCNS	0.21	0.19	0.61	0.12	0.61	0.26	-0.03	0.31	-0.02	-0.08	0.09	0.64	0.53	0.35	0.11	0.47	0.45	0.72	0.44	1

EDA

All plots used in Exloratory Data Analysis used to form intial hypothesis are save in the pdf

eda.pdf

Missing variables imputation



Partial Plots

The file contains partial plots of top 15 predictors in terms of decreasing covariance explanation wrt to logarithmic transformation of Energy and Water consumption

partial_plots.pdf

These are the interaction plots between top 4 predictors

interactions_plots.pdf

References

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. http://doi.org/10.1016/j.amc.2006.08.093

Melek Yalcintas1, U. A. O. (2007). An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. *International Journal of Energy Research*, *31*(August 2007), 135–147. http://doi.org/10.1002/er

Patrick J. Miller, Gitta H. Lubke, Daniel B. McArtor, C. S. Bergeman (2016). Finding structure in data using multivariate tree boosting. https://arxiv.org/ftp/arxiv/papers/1511/1511.02025.pdf