Prediction of Electricity Consumption

Yung Ching Chen

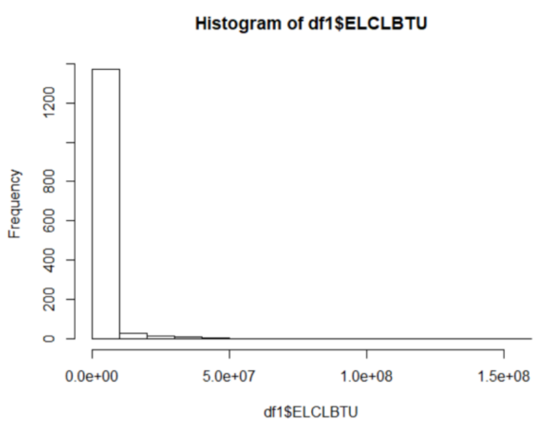
1. Introduction to Dataset

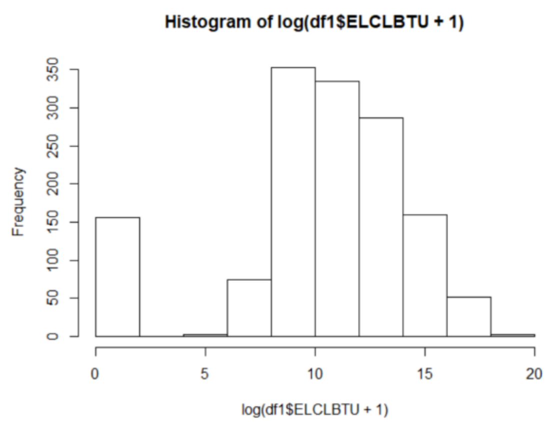
1.1 Response Variable and Transformation

The response variable we want to predict here is ELCLBTU, the electricity cooling use in thousand Btu.

Response variable	Min	Median	Mean	Max
ELCLBTU	0	41094	1347602	154764554

We find the distribution of ELCLBTU is extremely skewed, thus we do log transformation and get a new response variable log(ELCLBTU+1) that is more close to a normal distribution.





Response variable	Min	Median	Mean	Max
log(ELCLBTU+1)	0	10.624	10.169	18.857

1.2 Data Cleaning

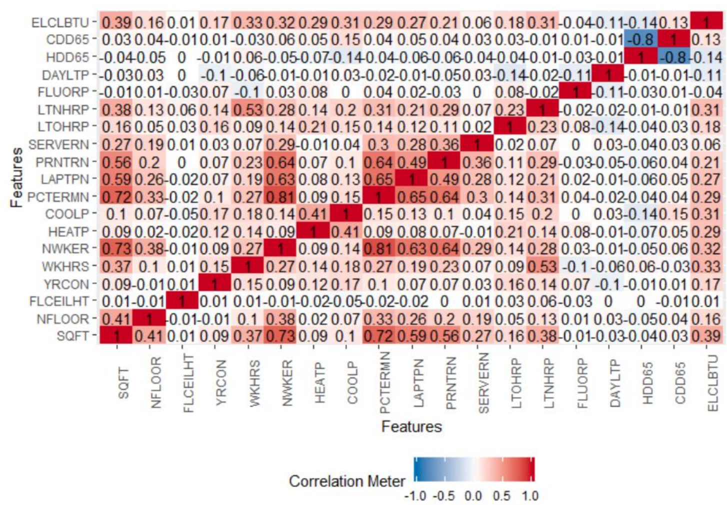
- Select rows that are in Midwest region (region = 2).
- Remove imputation flags, statistical weights and energy consumption columns except ELCLBTU.
- Remove columns that contain more than 30% of NA values.
- Remove rows that contain more than 50% of NA values.
- Conduct log transformation on response variable ELCLBTU.

- Transform continuous variables into numeric type and categorical variables into factor type.
- Remove factors (FREESTN, ELUSED, MFUSED) that contain only one level.
- Impute NA values using predictors with missForest package

1.3 Correlation Analysis

- PCTERMN is highly positively correlated to NWKER (r = 0.81)
- HDD65 is highly negatively correlated to CDD65 (r = -0.8)

Thus, we delete PCTERMN and CDD65.



1.4 Splitting Dataset

80% of data will be randomly chosen to be the training set, and the rest 20% will be the test set.

2. Feature Selection

We select important variables from the random forest models. To be more specific, the predictors will be chosen only if %IncMSE is larger or equal to 5. Totally 28 predictors are chosen, and the below table contains more information.

Column Name	Definition	Type
ELCOOL	Electricity used for cooling	Factor
SQFT	Square footage	Numeric
PBAPLUS	More specific building activity	Factor
COOL	Energy used for cooling	Factor
SQFTC	Square footage category	Factor
NWKER	Number of employees	Numeric
COOLP	Percent cooled	Numeric

PBA	Principal building activity	Factor
MAINCL	Main cooling equipment	Factor
HDD65	Heating degree days (base 65)	Numeric
NWKERC	Number of employees category	Factor
WKHRSC	Weekly hours category	Factor
PCTRMC	Number of computers category	Factor
CHILLR	Central chillers inside the building	Factor
PRNTRN	Number of printers	Numeric
HWRDHT	How reduce heating	Factor
WKHRS	Total hours open per week	Numeric
CWUSED	District chilled water used	Factor
NFLOOR	Number of floors	Numeric
PKGCL	Packaged A/C units	Factor
RFGICE	Commercial ice makers	Factor
BOILER	Boilers inside the building	Factor
CHWT	District chilled water piped in	Factor
MAINT	Regular HVAC maintenance	Factor
LAPTPN	Number of laptops	Numeric
SCHED	Light scheduling	Factor
OPNWE	Open on weekend	Factor
EMCS	Building automation system	Factor

3. Modeling Building

- Linear regression with 10-fold cross validation
- Stepwise linear regression
- Random forest *(tuned)*
- GAM
- MARS *(tuned)*
- BART *(tuned)*
- SVM *(tuned)*

The below table are the details related to tuning processes in each model.

Model	Hyperparameters	Scope	Best Value
Random forest	mtry	[8,9…28]	28
MARS	nprune	[5,10,15,20]	20
	degree	[1,2,3]	2
BART	tune num_tree_cvs	[50, 200]	200
	k_cvs	[2,3,5]	5
	nu	[3,10]	10

	q	[0,9, 0.99, 0.75]	0.75
SVM	epsilon	[0.1,0.2…,1]	0.1
	cose	2 ^ [2,3…,9]	16

4. Model Selection

After we build different models, we would like to evaluate their performance using RMSE and MAE both on in-sample dataset and out-of-sample dataset.

Model	In-Sample RMSE	Out-of-Sample RMSE	In-Sample MAE	Out-of-Sample MAE
Null model	4.23	4.39	3.05	3.11
Linear regression	0.80	6.13	0.61	4.49
Stepwise	0.82	6.13	0.62	4.49
Random forest	0.90	6.01	0.66	4.36
Random forest (tuned)	0.86	6.08	0.62	4.41
GAM	0.81	6.13	0.61	4.49
MARS	0.87	6.13	0.67	4.48
MARS (tuned)	0.81	6.13	0.62	4.49
BART	0.51	0.88	0.39	0.64
BART (tuned)	0.64	0.84	0.48	0.62
SVM	1.10	5.91	0.71	4.35
SVM (tuned)	0.58	6.15	0.45	4.51

5. Final Model and Inference

After calculating the errors, we notice that BART that has been tuned has the best performance in out-of-sample dataset. Thus, we choose this tuned BART as our final model. The best hyperparameters in this model is shown in the below table.

Parameter	num_trees	k	nu	q
Value	200	5	10	0.75

Since BART is a very flexible model, it's hard to give clear numeric inference. However, we can use var_selection_by_permute, important_vars_global_max_names, important_vars_local_names functions in bartMachine package to find the most important predictors. These include SQFT, COOLP, ELCOOL_1, ELCOOL_2, NWKER, PBA_13, SQFTC_4, SQFTC_2, PBA_2, PBA_8, NWKERC_2, PBA_15, PBAPLUS_19, PBAPLUS_32, PBA_16, PBAPLUS_1, PBAPLUS_2. Among them, SQFT, COOLP, ELCOOL_1, ELCOOL_2, NWKER are most significant. Thus, we may make the below conclusions based on the model.

- The bigger the square footage, percent cooled or number of employees of a building, the higher the electricity used for cooling.
- Electricity used for cooling is a significant factor to predict the electricity used for cooling of a building.