Dataset Description and Summary Statistics

The dataset for implementing the gradient descent algorithm for linear regression is downloaded from UCI Machine Learning repository. The link for downloading the dataset is appliances+energy+prediction.

Dataset consists of 19735 observations on a total of 29 variables. The variables are Appliances and lights energy consumption in the house, 9 temperatures and 9 relative humidity in 9 different locations in the house, 6 weather variables such as outside temperature, outside humidity, Pressure, Windspeed, Visibility and Dewpoint. We also have date column, and two random variables that are being ignored in this model. Additional information about the dataset can be found out in the above mentioned link. The dependent variable for linear regression model is Appliances energy consumption based on the remaining predictors. There are no missing values in any column, so no missing value imputation is required on this dataset.

Train/Test Split, Classification, Standardization

The dataset is split into train and test sets using 70-30 split percentage. The number of observations in train and test sets are 13814 and 5921 respectively.

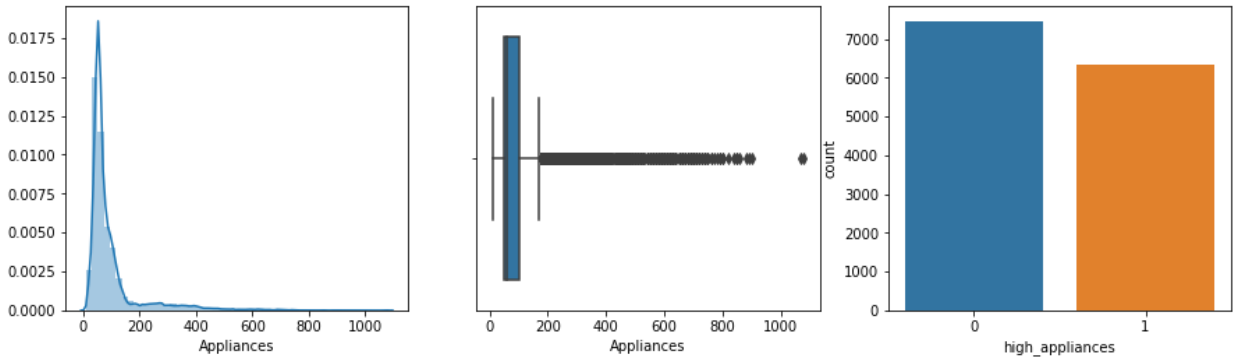
A new target variable, *high appliances* is created using Appliances median energy consumption as cutoff. Houses with appliances energy consumption more than median are denoted as 1 and the remaining as 0.

The independent variables in train set are standardized using mean and standard deviation of respective features from the train set. The features in test set are also standardized using the same train set mean and standard deviation to avoid data leakage from test set into train set which might give overfit model.

Data Exploration – Target Variable

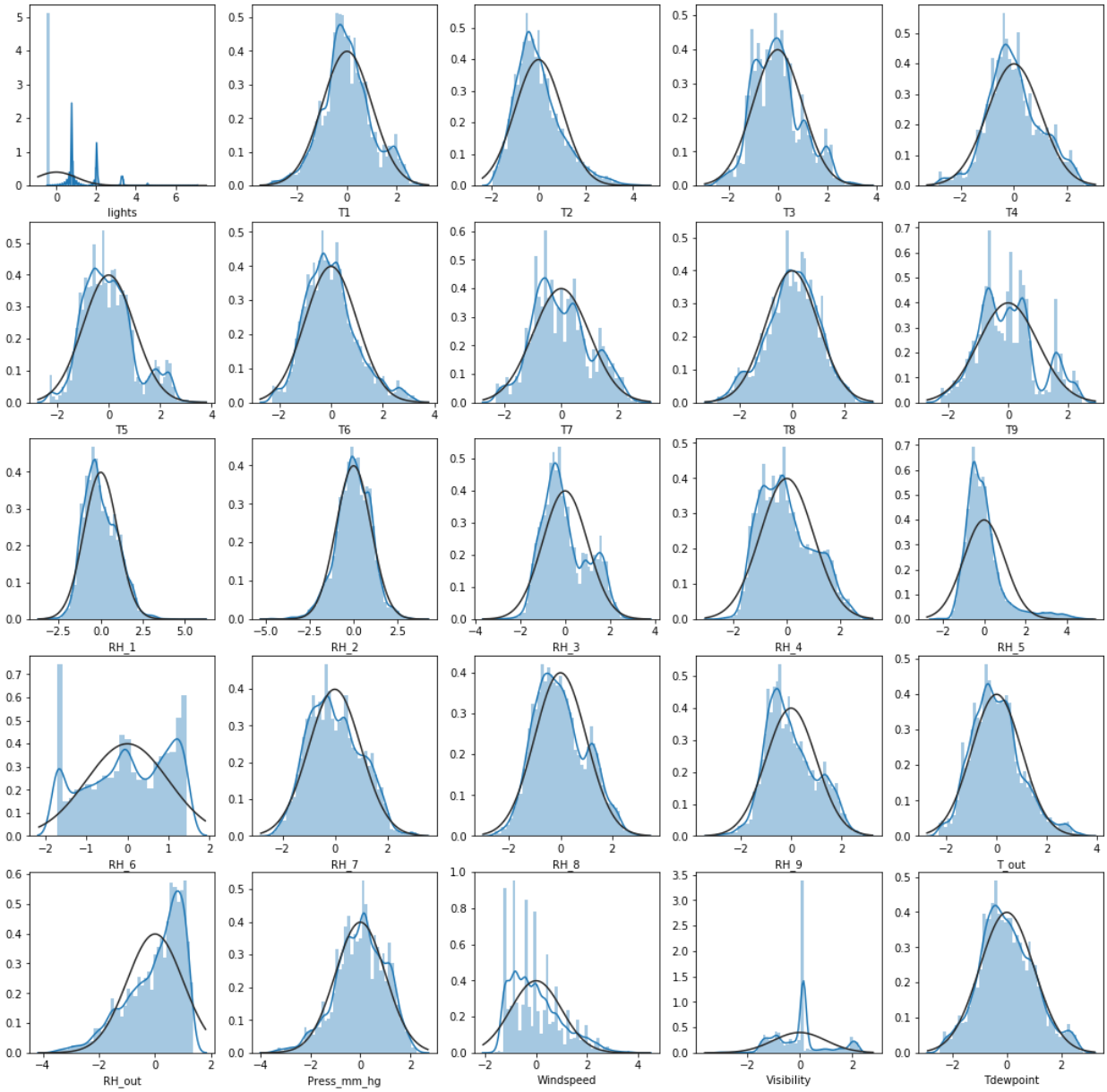
The target variable is positively skewed with few outliers. But they are ignored since there is no strong evidence on the outliers being clerical errors. The data is collected using sensors and faulty data could result only from sensor malfunction. Since, we don't have information on this, the outliers are taken as legitimate data and are included in the model. Moreover, they contribute only 2.7% of the observations.

For the category variable in train, there are 6351 houses with high appliances energy consumption (greater than median) and 7463 houses with low appliance energy consumption (less or equal to median).



Data Exploration - Predictor Variables

Below are distribution charts of 25 features. Except for RH_6, all others roughly fit normal distribution

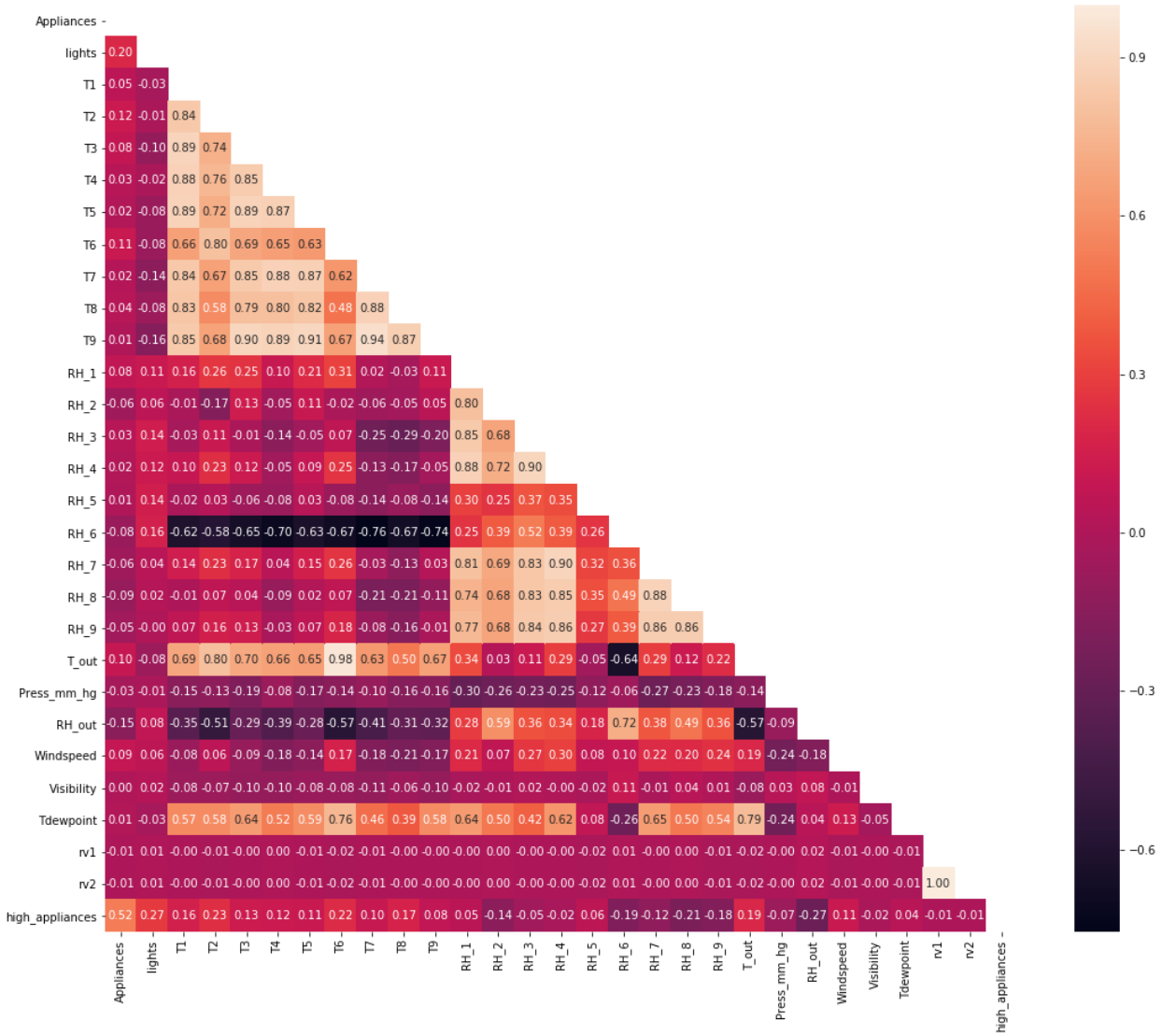


Feature Selection

Features for different experiments are selected using the correlation heatmap presented in next page.
15+ features models - 17 features are shortlisted that have highest correlation with the target variable.
10 random features model - 10 random features are selected form all available 25 predictor variables.
10 best features model - 10 features are shortlisted that have highest correlation with target variable.

RH_4 has 0.90 correlation with RH_3,RH_7 which have better correlation with target than RH_4.
T9 has more than 0.90 correlation with T3,T5,T7 which have better correlation with target than T9
T_out has 0.98 correlation with T6 which has better correlation with the target variable than T_out.

Correlation Heatmap



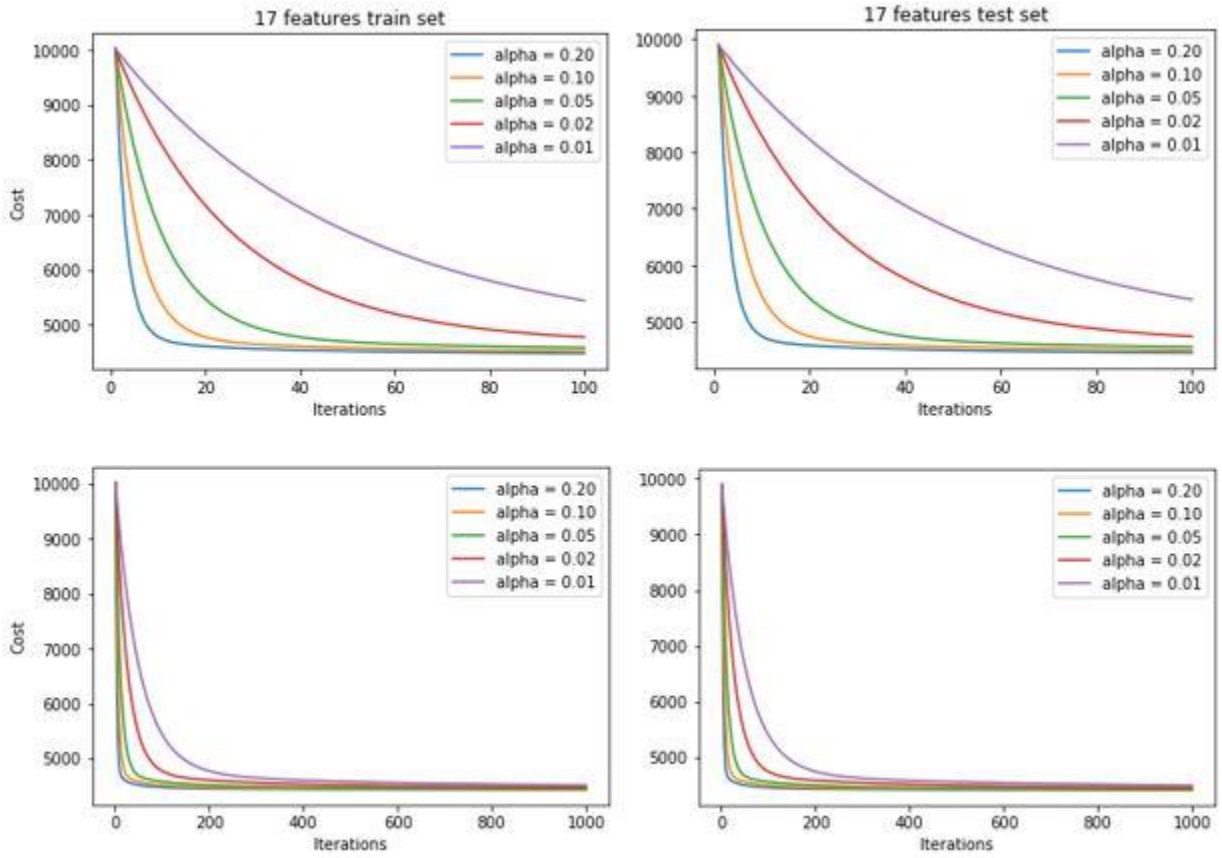
Models Specification

Model 1 (Experiment 1) - linear (target = *Appliances*) and logistic (target = *high_appliances*) regression
target = β_0 + β_{lights}*lights + β_{T1}*T1 + β_{T2}*T2 + β_{T3}*T3 + β_{T4}*T4 + β_{T6}*T6 + β_{T8}*T8 + β_{RH1}*RH1 + β_{RH2}*RH2 + β_{RH3}*RH3 + β_{RH6}*RH6 + β_{RH7}*RH7 + β_{RH8}*RH8 + β_{RH9}*RH9 + β_{Press}*Press_mm_hg + β_{RH_out}*RH_out + $\beta_{Windspeed}$*Windspeed

Model 2 (Experiment 3)
target = β_0 + β_{T1}*T1 + β_{T3}*T3 + β_{T5}*T5 + β_{T8}*T8 + β_{RH3}*RH3 + β_{RH5}*RH5 + β_{Press}*Press_mm_hg + β_{RH_out}*RH_out + $\beta_{Visibility}$*Visibility + $\beta_{Tdewpoint}$*Tdewpoint

Model 3 (Experiment 4)
target = β_0 + β_{lights}*lights + β_{T2}*T2 + β_{T3}*T3 + β_{T6}*T6 + β_{RH1}*RH1 + β_{RH2}*RH2 + β_{RH6}*RH6 + β_{RH8}*RH8 + β_{RH_out}*RH_out + $\beta_{Windspeed}$*Windspeed

Experiment 1a - Tuning alpha parameter with gradient descent for linear regression with convergence threshold of 0.0001. Following graphs shows the decrease in cost function against number of iterations.



We can observe that as value of alpha decreases, the algorithm needs more iterations for convergence. Following table shows the actual number of iterations needed for above alphas with threshold of 10^{-16}.

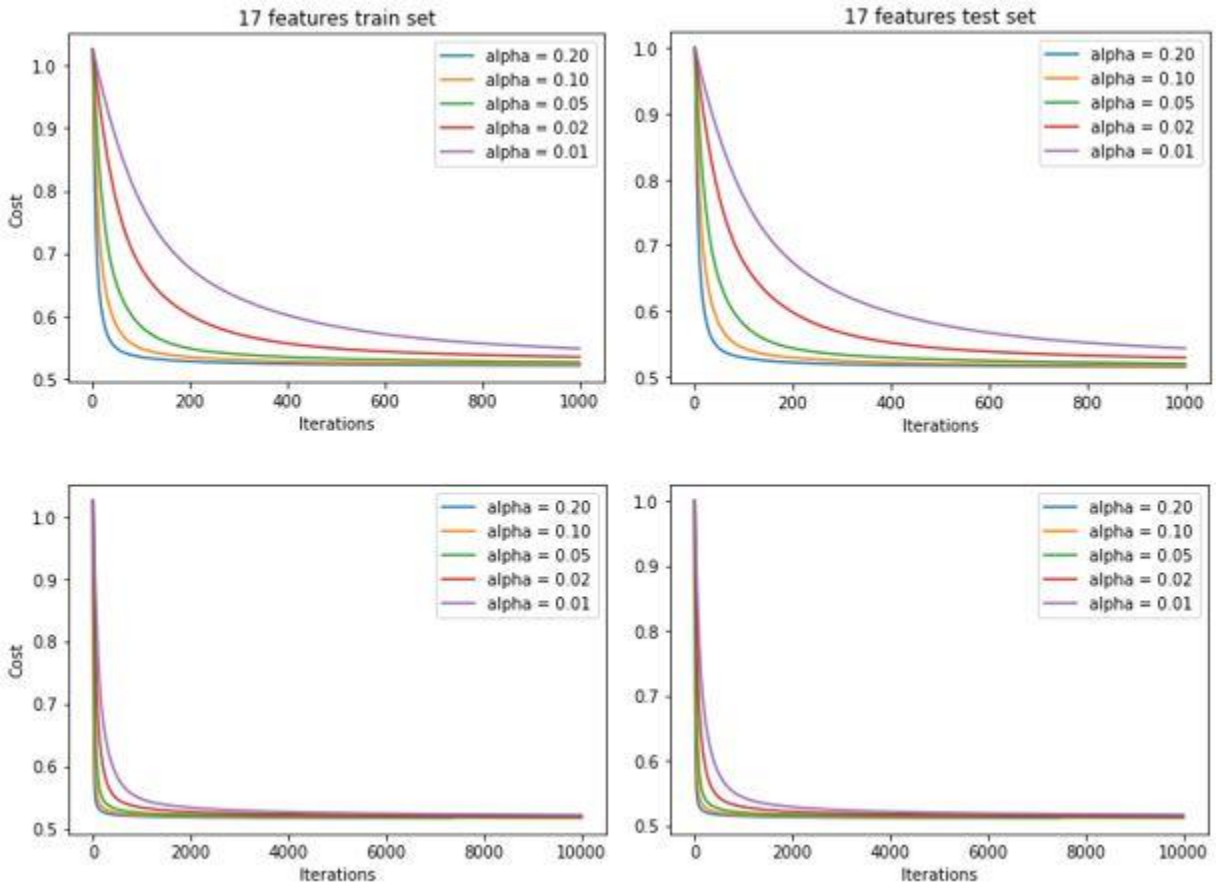
Learning Rate	Convergence Threshold	Convergence Time	Iterations	Cost after convergence
0.20	$\mathbf{10^{-16}}$	**7 seconds**	**3999**	**4440.6617709813245**
0.10	10^{-16}	14 seconds	7773	4440.6617709815050
0.05	10^{-16}	26 seconds	14978	4440.6617709820380
0.02	10^{-16}	67 seconds	35874	4440.6617709836500
0.01	10^{-16}	133 seconds	69513	4440.6617709863895

For alpha = 0.2 and convergence threshold = 10^{-16}, below beta parameters came from gradient descent

β_0	= 98.0129	β_{T6}	= 12.5884	β_{RH7}	= -9.1124
β_{lights}	= 18.5415	β_{T8}	= 7.1482	β_{RH8}	= -23.3118
β_{T1}	= 1.0971	β_{RH1}	= 63.3958	β_{RH9}	= -6.7171
β_{T2}	= -40.1456	β_{RH2}	= -62.2118	β_{Press}	= 1.6765
β_{T3}	= 43.7432	β_{RH3}	= 20.4086	β_{RHout}	= 0.8448
β_{T4}	= -19.2265	β_{RH6}	= 8.6189	$\beta_{Windspeed}$	= 4.2329

Coefficients from gradient descent are accurate to the closed form linear regression model coefficients. With linear model package, R^2-statisitc is found as 15.6% and T2, Pressure, RH are insignificant features.

Experiment 1b - Tuning alpha parameter with gradient descent for logistic regression with convergence threshold of 10^{-8}. Following graphs shows the decrease in cost function against number of iterations.



We can observe that as value of alpha decreases, the algorithm needs more iterations for convergence. Following table shows the actual number of iterations needed for above alphas with threshold of 10^{-16}.

Learning Rate	Convergence Threshold	Convergence Time	Iterations	Cost after convergence
0.20	$\mathbf{10^{-16}}$	**48 seconds**	**29004**	**0.5175267603268712**
0.10	10^{-16}	83 seconds	55549	0.5175267603272693
0.05	10^{-16}	179 seconds	109001	0.5175267603276069
0.02	10^{-16}	563 seconds	258778	0.5175267603296202
0.01	10^{-16}	1325 seconds	502194	0.5175267603322398

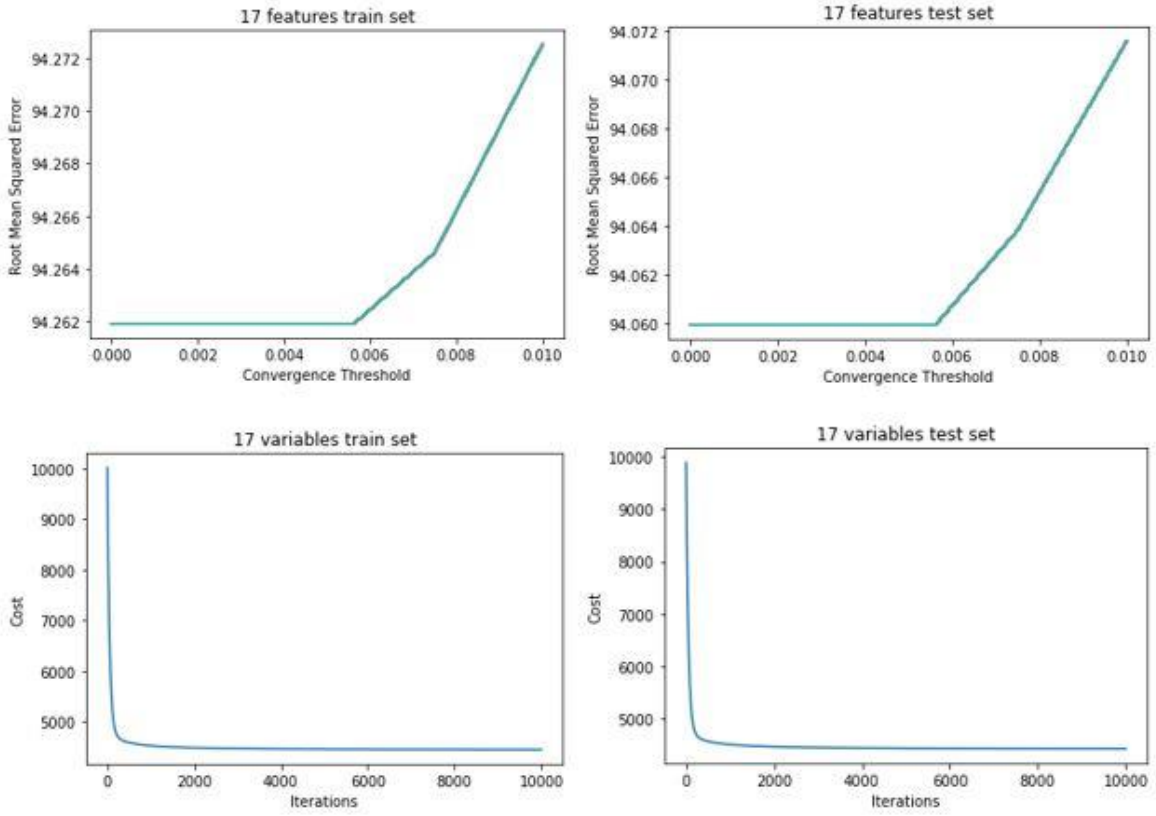
For alpha = 0.2 and convergence threshold = 10^{-16}, below beta parameters came from gradient descent

β_0	= -0.15078	β_{T6}	= 0.56904	β_{RH7}	= -0.03968
β_{lights}	= 0.68980	β_{T8}	= 0.76855	β_{RH8}	= -0.78382
β_{T1}	= 0.47667	β_{RH1}	= 2.18618	β_{RH9}	= -0.72758
β_{T2}	= -0.70678	β_{RH2}	= -1.43367	β_{Press}	= 0.01965
β_{T3}	= 0.35108	β_{RH3}	= 0.08040	β_{RHout}	= -0.11551
β_{T4}	= -0.85176	β_{RH6}	= 0.54689	$\beta_{Windspeed}$	= 0.20180

Coefficients from gradient descent are accurate to the closed form logistic regression model coefficients.

Experiment 2

Following graphs shows the Root Mean Squared Error as a function of different convergence thresholds



Above graphs show cost function for threshold at 0.005 and its relationship with number of iterations.

Experiment 3

Following table shows the errors and RMSE obtained from 17 selected features and 10 random features.

Model	Iterations	Cost	RMSE
17 features train	**3999**	**4440.66**	**94.24**
10 random train	781	4972.68	99.73
17 features test	**3845**	**4421.81**	**94.04**
10 random test	807	4934.00	99.34

Model with 17 features has lower cost and error statistics, however it takes more iterations to converge.

Experiment 4

Following table shows the errors and RMSE obtained from 17 selected features and 10 best features.

Model	Iterations	Cost	RMSE
17 features train	**3999**	**4440.66**	**94.24**
10 best train	1938	4504.05	94.91
17 features test	**3845**	**4421.81**	**94.04**
10 best test	1900	4470.48	94.56

Model with 17 features has lower cost and error statistics, however it takes more iterations to converge.