Indian Institute of Technology Madras

M.Tech (Computer Science and Engineering)

CS6464 : Concepts in Statistical Learning Theory

Software Assignment 2

SUBMITTED BY:
CS18M038, Ojas Mehta
CS18M052, Avani Shukla

Team Number: 5

1 Task

Predict house prices given training and test house data of 20- dimensional features and comparing the performance of various regression methods.
Allotted regression methods are:

1. Ridge Regression
2. Backward Stepwise Regression
3. Kernel Regression

2 Result

- Data Preprocessing: Removed Date which was a string(Non-numeric values) and removed ID for each row which clearly was insignificant as a predictor.

- Data Analysis: Reponse variable is "Price"
  predictors like sqft_living, sqft_above, grade, bathrooms, sqft_living15 are very highly correlated with Response variable and seems most important
  Predictors like view, bedrooms and lat are moderately correlated with Response variable.
  Predictors like view, bedrooms and lat are moderately correlated with Response variable.
  Predictors like sqft_lot, condition yr_built, zipcode and long are very less correlated with Response variable and can almost be considered independent to it.

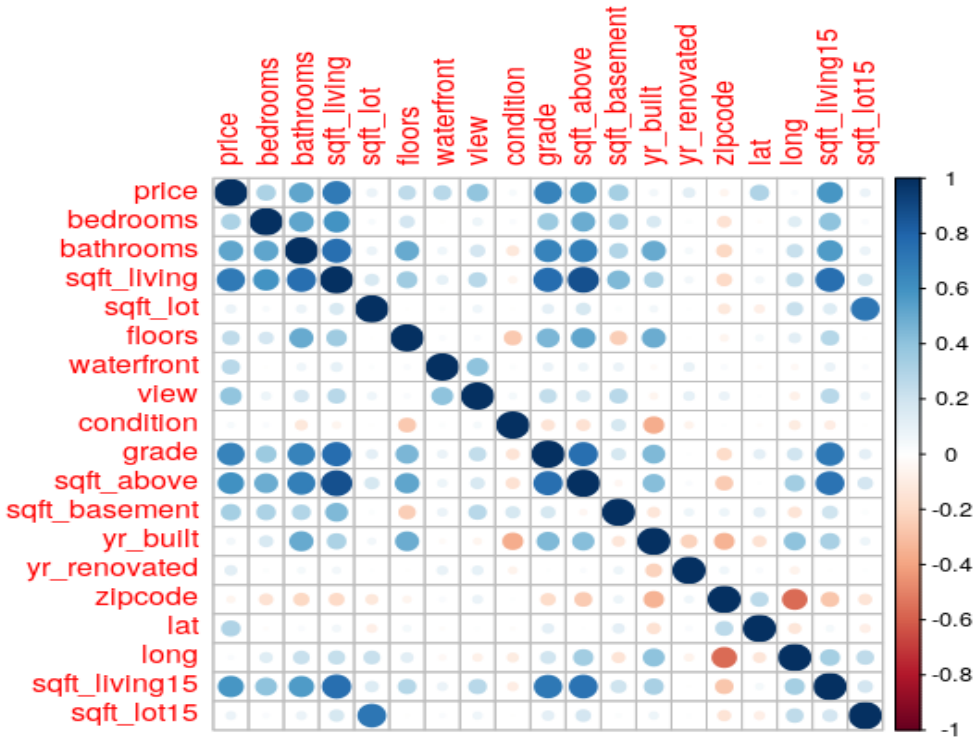

Figure 1: Correlation of predictors with response variable(price)

Feature pair like (sqft_living,sqft_above), (bathrooms, sqft_living), (sqft_living,sqft_living15) and (grade,sqft_living) have very high correlation, So, any one of them can be selected as a predictor to reduce the number of features.

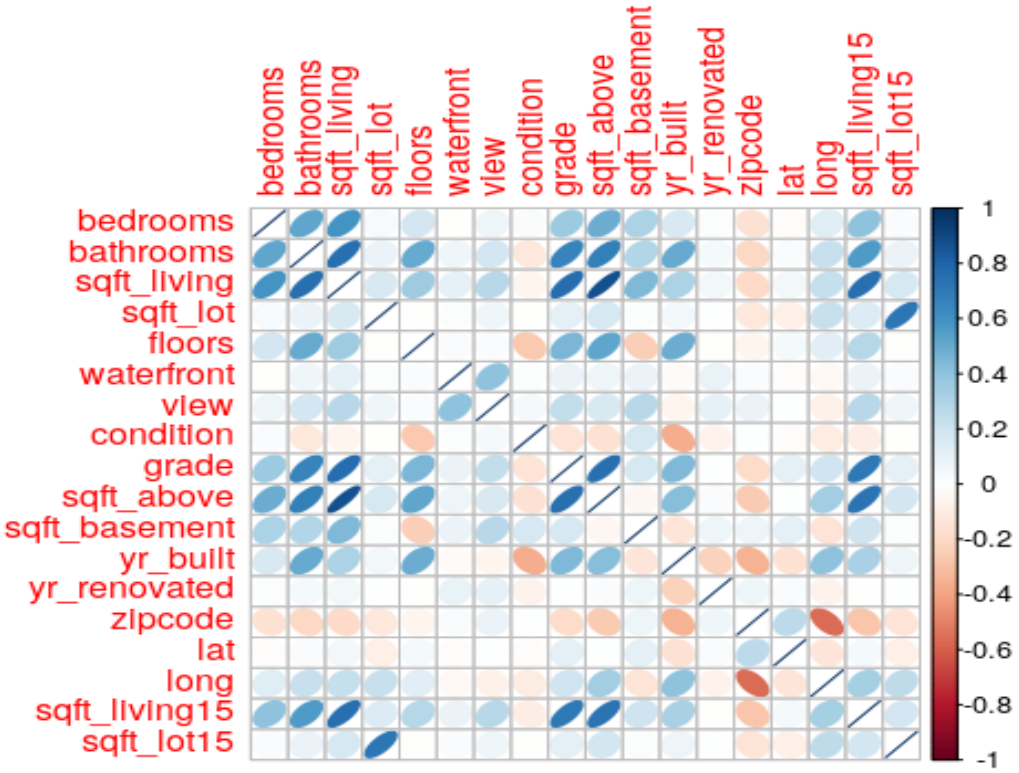

Figure 2: Correlation of predictors

- Coefficients/Regression Weights Features corresponding to the largest magnitude of the coefficient/weights are taken as the most important as the response is most sensitive to the changes in those predictors.

1. Ridge Regression

bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
-36060.282	32177.854	83010.041	6554.545	2654.901	52818.336	38237.428	16434.043
grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
115847.125	77112.525	28632.189	-78726.338	7868.407	-32452.070	83688.378	-30317.634
sqft_living15	sqft_lot15						
14055.796	-10626.977						

Figure 3: Coefficients

grade and sqft_living are the most significant predictors.

2. Backward Stepwise Regression

(Intercept)	539366.63
bedrooms	-41830.95
bathrooms	17636.21
sqft_living	183485.82
sqft_lot	.
view	65052.31
condition	33805.44
sqft_above	48168.31
yr_renovated	23768.84
lat	93943.94
long	-40992.09
sqft_living15	41938.69

Figure 4: Coefficients

Backward Stepwise Regression has selected 11 predictors and out of them sqft_living is the most significant predictors.

3. Kernel Regression
   Kernel regression was done using gassian kernel.

bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
0.01605784	0.01671253	0.11731738	0.31089061	0.02040010	0.40317388	0.02595653	0.03332538
grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
0.16515847	0.15851675	-0.04553558	0.08308993	-0.18138712	-0.12099274	0.01878245	-0.27838693
sqft_living15	sqft_lot15						
0.05037159	-0.04484218						

Figure 5: Coefficients

sqft_lot and waterfront are the most significant predictors.

- Plot of Coefficient Profile
  To get the top 5 interesting features, calculate the variance of each feature over iterations and select the 5 features with highest variance.

  1. Ridge Regression

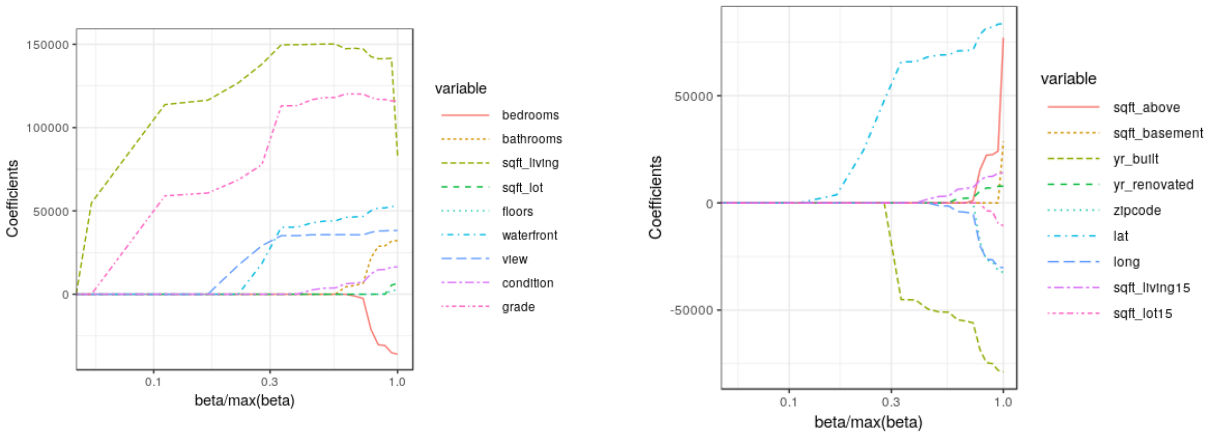

Figure 6: Plot of Coefficient profile with all predictors

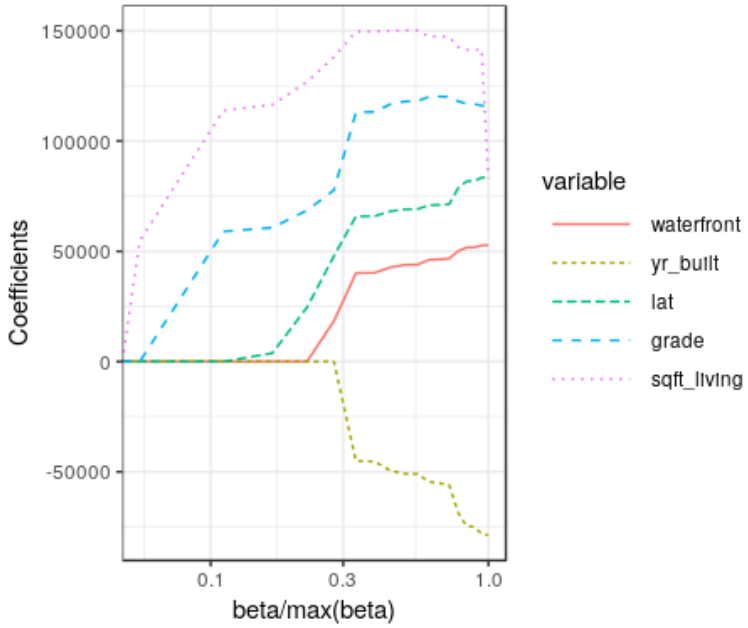

Figure 7: Plot of Coefficient profile of top 5 interesting features

2. Backward Stepwise Regression

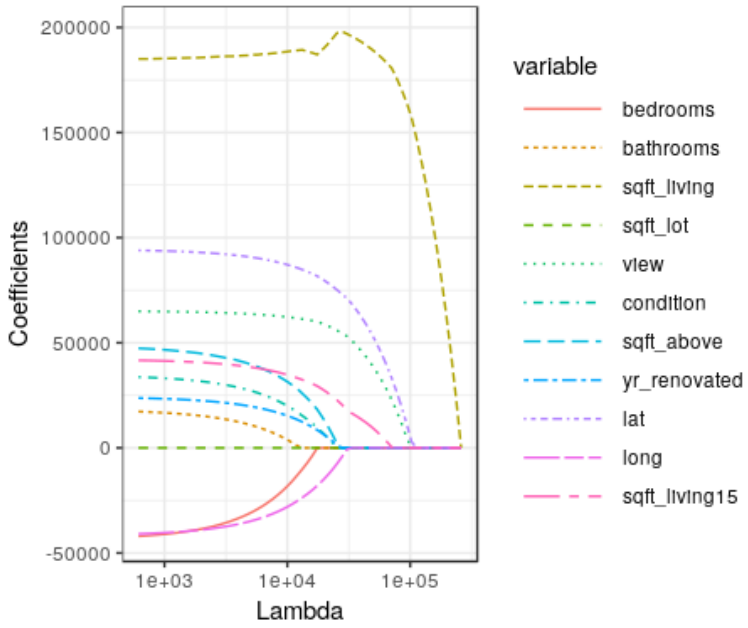

Figure 8: Plot of Coefficient profile with selected 11 predictors

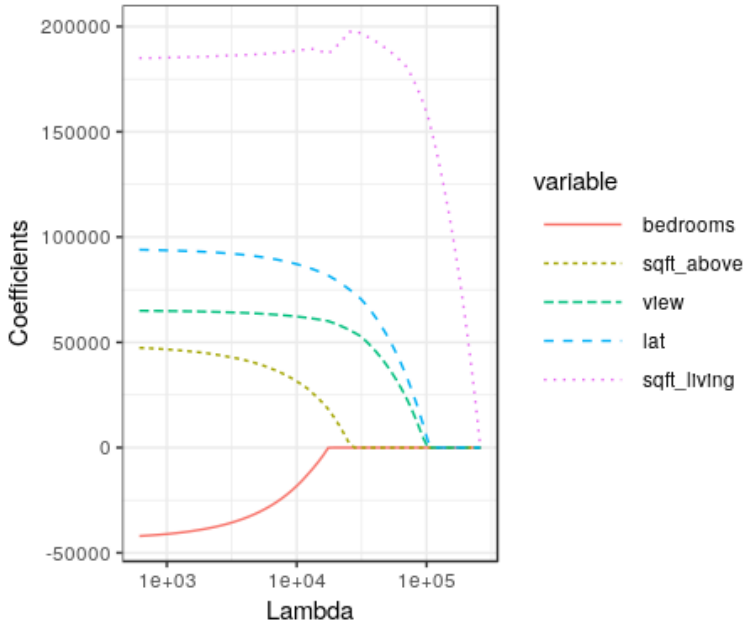

Figure 9: Plot of Coefficient profile of top 5 interesting features

3. Kernel Regression

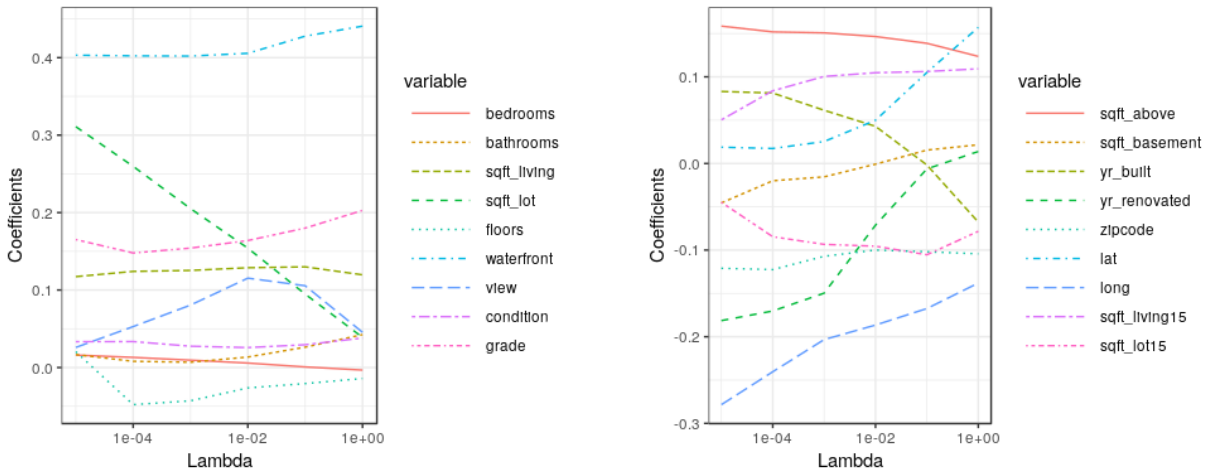

Figure 10: Plot of Coefficient profile with all predictors

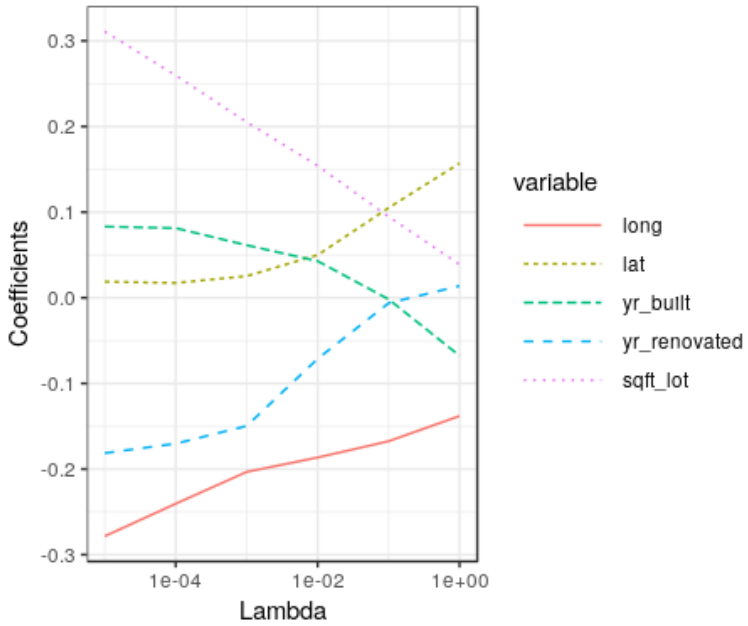

Figure 11: Plot of Coefficient profile of top 5 interesting features

- RSS/RMSE of Model

$$RSS = \sum_{N}(actual\_price - predicted\_price)$$

$$RMSE = \sqrt{\frac{1}{N}RSS}$$

  - Train Data:

	Ridge($\lambda = 0.001$)	Backward Stepwise	Kernel(Guassian)
RSS	$6.727\,335 \times 10^{15}$	$6.570\,453 \times 10^{15}$	8879.584
RMSE	622080.6	614784.3	42.14163

– Test Data:

	Ridge($\lambda = 0.001$)	Backward Stepwise	Kernel(Guassian)
RSS	$1.636\,735 \times 10^{15}$	$1.602\,115 \times 10^{15}$	2995.31
RMSE	622114.5	615499.8	24.47574

- Conclusion:
  - RSS is calculated for 10,000 over test data, so summation over this many data point gives large value of RSS.
  - From the above two tables, RSS value is minimum for kernel Regression with guassian kernel method.
  - Clearly we can observe that linear method for regression (Ridge and Backward Stepwise) is not successful for predicting house price whereas Guassian kernel regression can predict house price close to actual price.