Submitted by:
Kiran Kumar Valluru (APFE21709855)

Part II - Assignment-based Subjective Questions

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

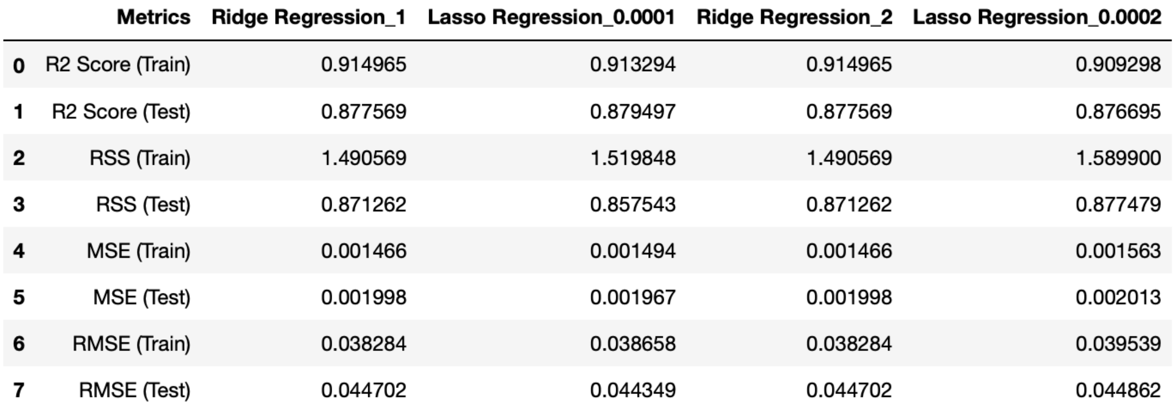
Optimal value of alpha for Ridge and Lasso Regression is below:

- Alpha value for Ridge Regression: 0.6
- Alpha value for Lasso Regression: 0.0001

With these alphas the R2 score for the models is approximately 0.87.

After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.87 with minor reduction in both Train and Test R2 score and there is a small change in the co-efficient values. The new model is created and demonstrated in the Jupyter notebook.

Comparison of metrics for Ridge and Lasso Regression with Original and Double the alpha:

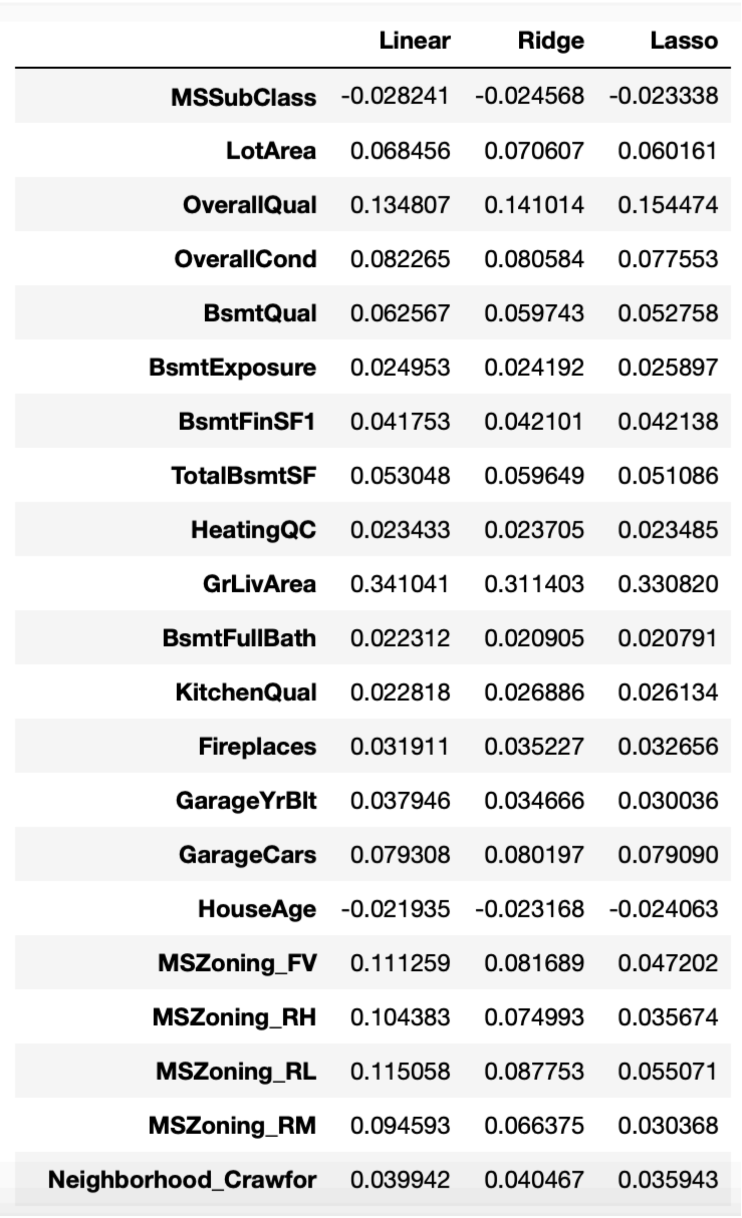
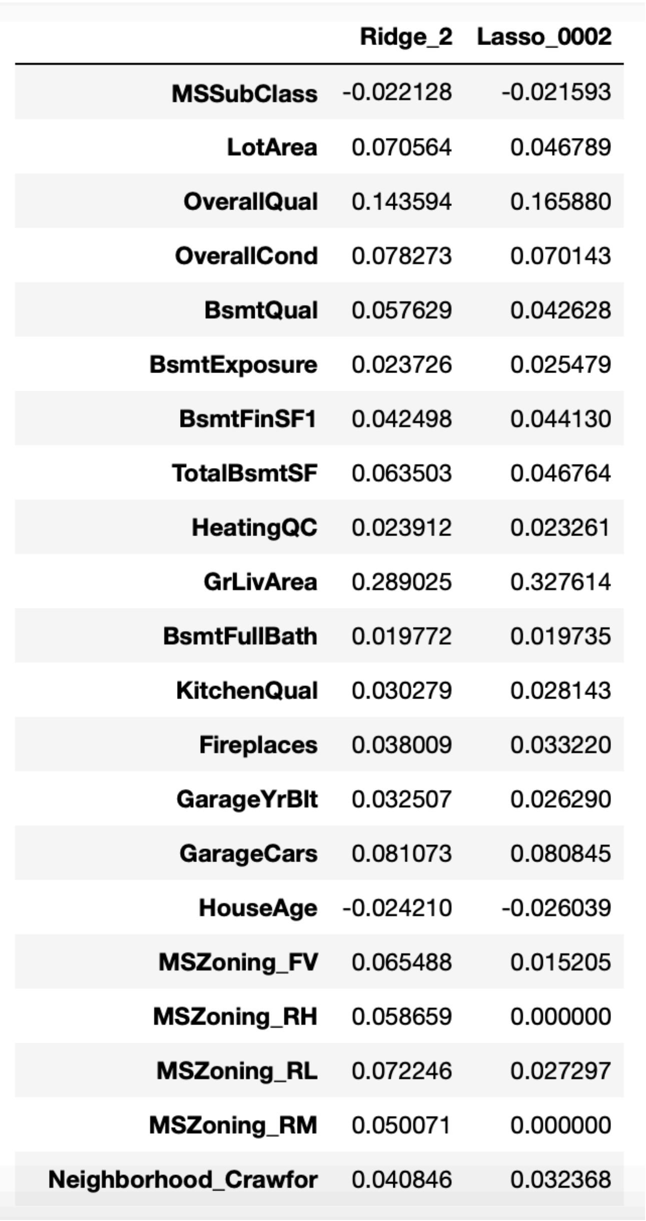


	Metrics	Ridge Regression_1	Lasso Regression_0.0001	Ridge Regression_2	Lasso Regression_0.0002
0	R2 Score (Train)	0.914965	0.913294	0.914965	0.909298
1	R2 Score (Test)	0.877569	0.879497	0.877569	0.876695
2	RSS (Train)	1.490569	1.519848	1.490569	1.589900
3	RSS (Test)	0.871262	0.857543	0.871262	0.877479
4	MSE (Train)	0.001466	0.001494	0.001466	0.001563
5	MSE (Test)	0.001998	0.001967	0.001998	0.002013
6	RMSE (Train)	0.038284	0.038658	0.038284	0.039539
7	RMSE (Test)	0.044702	0.044349	0.044702	0.044862

Ridge: Train R2 score reduces little bit from 91.63% to 91.50% and Test R2 score also reduce from 87.96% to 87.76%.

Lasso: Train R2 score reduces little bit from 91.33% to 90.93% and Test R2 score also reduce from 87.95% to 87.67%.

Below are the changes in the co-efficients:

<table>
<tr><th>Ridge and Lasso Coefficients (Original)</th><th>Ridge and Lasso Coefficients (after doubling the Alpha)</th></tr>
<tr><td>

<table>
<tr><th></th><th>Linear</th><th>Ridge</th><th>Lasso</th></tr>
<tr><td>MSSubClass</td><td>-0.028241</td><td>-0.024568</td><td>-0.023338</td></tr>
<tr><td>LotArea</td><td>0.068456</td><td>0.070607</td><td>0.060161</td></tr>
<tr><td>OverallQual</td><td>0.134807</td><td>0.141014</td><td>0.154474</td></tr>
<tr><td>OverallCond</td><td>0.082265</td><td>0.080584</td><td>0.077553</td></tr>
<tr><td>BsmtQual</td><td>0.062567</td><td>0.059743</td><td>0.052758</td></tr>
<tr><td>BsmtExposure</td><td>0.024953</td><td>0.024192</td><td>0.025897</td></tr>
<tr><td>BsmtFinSF1</td><td>0.041753</td><td>0.042101</td><td>0.042138</td></tr>
<tr><td>TotalBsmtSF</td><td>0.053048</td><td>0.059649</td><td>0.051086</td></tr>
<tr><td>HeatingQC</td><td>0.023433</td><td>0.023705</td><td>0.023485</td></tr>
<tr><td>GrLivArea</td><td>0.341041</td><td>0.311403</td><td>0.330820</td></tr>
<tr><td>BsmtFullBath</td><td>0.022312</td><td>0.020905</td><td>0.020791</td></tr>
<tr><td>KitchenQual</td><td>0.022818</td><td>0.026886</td><td>0.026134</td></tr>
<tr><td>Fireplaces</td><td>0.031911</td><td>0.035227</td><td>0.032656</td></tr>
<tr><td>GarageYrBlt</td><td>0.037946</td><td>0.034666</td><td>0.030036</td></tr>
<tr><td>GarageCars</td><td>0.079308</td><td>0.080197</td><td>0.079090</td></tr>
<tr><td>HouseAge</td><td>-0.021935</td><td>-0.023168</td><td>-0.024063</td></tr>
<tr><td>MSZoning_FV</td><td>0.111259</td><td>0.081689</td><td>0.047202</td></tr>
<tr><td>MSZoning_RH</td><td>0.104383</td><td>0.074993</td><td>0.035674</td></tr>
<tr><td>MSZoning_RL</td><td>0.115058</td><td>0.087753</td><td>0.055071</td></tr>
<tr><td>MSZoning_RM</td><td>0.094593</td><td>0.066375</td><td>0.030368</td></tr>
<tr><td>Neighborhood_Crawfor</td><td>0.039942</td><td>0.040467</td><td>0.035943</td></tr>
</table>
</td><td>

<table>
<tr><th></th><th>Ridge_2</th><th>Lasso_0002</th></tr>
<tr><td>MSSubClass</td><td>-0.022128</td><td>-0.021593</td></tr>
<tr><td>LotArea</td><td>0.070564</td><td>0.046789</td></tr>
<tr><td>OverallQual</td><td>0.143594</td><td>0.165880</td></tr>
<tr><td>OverallCond</td><td>0.078273</td><td>0.070143</td></tr>
<tr><td>BsmtQual</td><td>0.057629</td><td>0.042628</td></tr>
<tr><td>BsmtExposure</td><td>0.023726</td><td>0.025479</td></tr>
<tr><td>BsmtFinSF1</td><td>0.042498</td><td>0.044130</td></tr>
<tr><td>TotalBsmtSF</td><td>0.063503</td><td>0.046764</td></tr>
<tr><td>HeatingQC</td><td>0.023912</td><td>0.023261</td></tr>
<tr><td>GrLivArea</td><td>0.289025</td><td>0.327614</td></tr>
<tr><td>BsmtFullBath</td><td>0.019772</td><td>0.019735</td></tr>
<tr><td>KitchenQual</td><td>0.030279</td><td>0.028143</td></tr>
<tr><td>Fireplaces</td><td>0.038009</td><td>0.033220</td></tr>
<tr><td>GarageYrBlt</td><td>0.032507</td><td>0.026290</td></tr>
<tr><td>GarageCars</td><td>0.081073</td><td>0.080845</td></tr>
<tr><td>HouseAge</td><td>-0.024210</td><td>-0.026039</td></tr>
<tr><td>MSZoning_FV</td><td>0.065488</td><td>0.015205</td></tr>
<tr><td>MSZoning_RH</td><td>0.058659</td><td>0.000000</td></tr>
<tr><td>MSZoning_RL</td><td>0.072246</td><td>0.027297</td></tr>
<tr><td>MSZoning_RM</td><td>0.050071</td><td>0.000000</td></tr>
<tr><td>Neighborhood_Crawfor</td><td>0.040846</td><td>0.032368</td></tr>
</table>
</td></tr>
</table>

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

The optimum lambda value in case of Ridge and Lasso is as follows:-

- Ridge – 0.6
- Lasso – 0.0001

The Mean Squared Error in case of Ridge and Lasso are:

- Ridge - 0.001965
- Lasso - 0.001967

The Mean Squared Error of both the models are almost same. R2 Score of Lasso is better than Ridge for Test Data.

Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), **Lasso** has a better edge over Ridge and should be used as the final model.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

The five most important predictor variables in the current lasso model is:-
1. GrLivArea: Above grade (ground) living area square feet
2. OverallQual: Rates the overall material and finish of the house
3. GarageCars: Size of garage in car capacity
4. OverallCond: Rates the overall condition of the house
5. LotArea: Lot size in square feet

Have build a Lasso model in the Jupyter notebook after removing these attributes from the dataset.

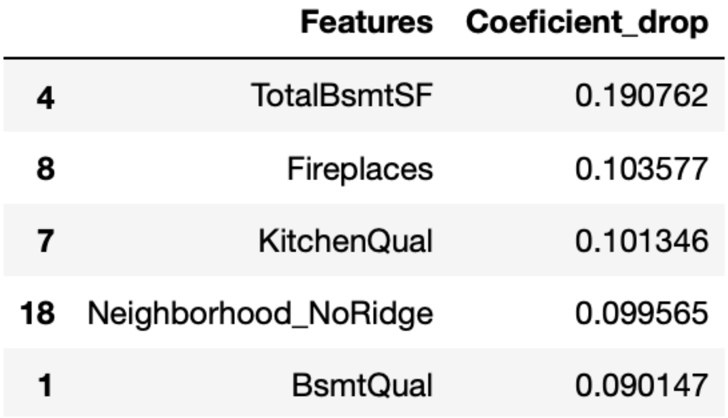
Train R2 score reduces drastically from 91.33% to 82.19% and Test R2 score also reduces from 87.95% to 73.65% after removal of Top 5 predictive variables.

The Mean Squared Error increases to 0.004300652262493751

The new Top 5 predictos after removing the Top 5 predictive variables are:-

1. TotalBsmtSF: Total square feet of basement area
2. FirePlaces: Number of fireplaces

3. KitchenQual: Kitchen quality
4. Neighborhood_NoRidge: Physical locations within Ames city limits - Northridge
5. BsmtQual: Evaluates the height of the basement



	Features	Coeficient_drop
4	TotalBsmtSF	0.190762
8	Fireplaces	0.103577
7	KitchenQual	0.101346
18	Neighborhood_NoRidge	0.099565
1	BsmtQual	0.090147

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

As Per, Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:

- Simpler models are usually more 'generic' and are more widely applicable.
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust.
  - Complex models tend to change wildly with changes in the training data set
  - Simple models have low variance, high bias and complex models have low bias, high variance
  - Simpler models make more errors in the training set. Complex models lead to overfitting - they work very well for the training samples, fail miserably when applied to other test samples.

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

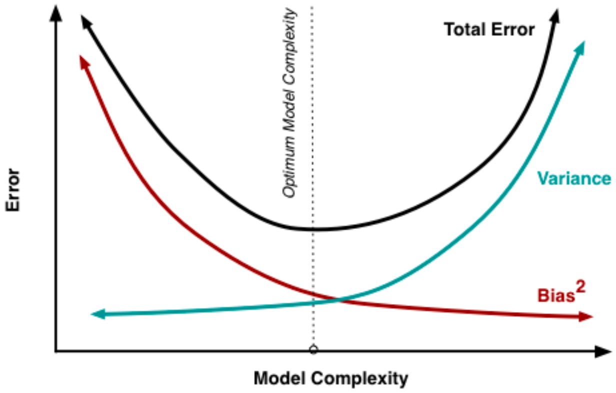
Also, Making a model simple leads to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.

Variance refers to the degree of changes in the model itself with respect to changes in the training data.

Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.



Below are the pointers learned during the course module as key items:

- Robust refers the model works for a broad range of inputs. If the model gets really good results at training time (it seems “more accurate”) but won’t generalize to out-of-sample data (i.e., it isn’t robust) then we call it overfitting.
- The model should be generalized so that the test accuracy is not lesser than the training score.
- Here in our case, based on all data and modelling both Ridge and Lasso performed good on Train and Test Data which shows our model with Alpha value "0.6" for Ridge and "0.0001" for Lasso is Robust and more Generalized model.
- Too much importance should not be given to the outliers so that the accuracy predicted by the model is high. But outlier’s analysis needs to be done and only those which are relevant to the dataset need to be retained and rest should be dropped.