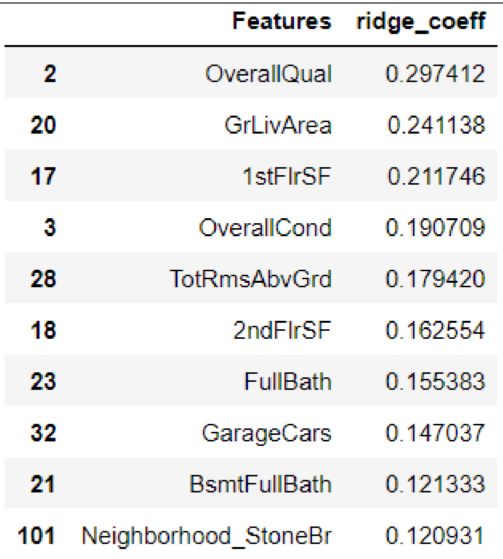
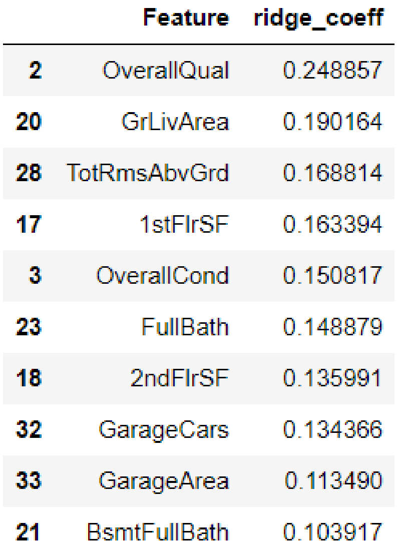
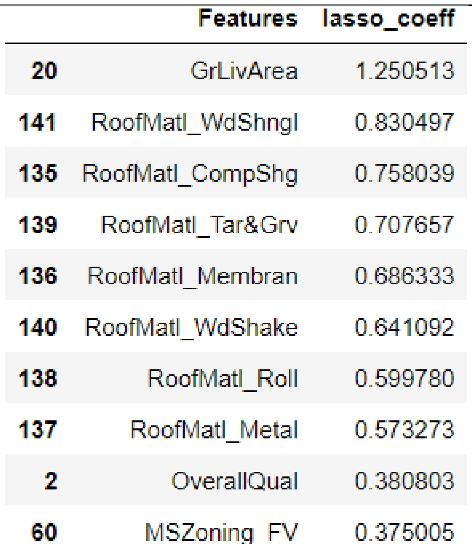
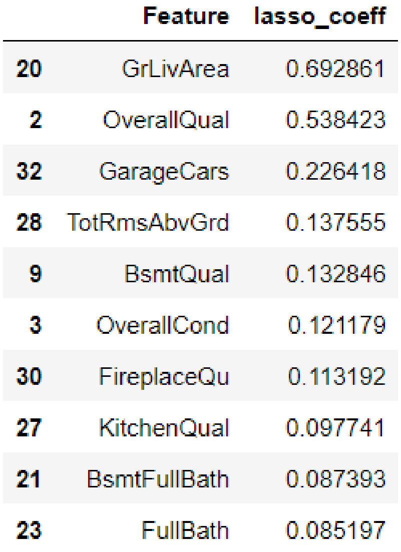
Question 1

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: The optimal value of ridge regression is 5 and for lasso regression it is 0.001.

With the increase in alpha, the penalty on the model increases which makes model too simple which increase the bias.

Below are how the top 10 features change in both the regression models after doubling the optimal alpha value

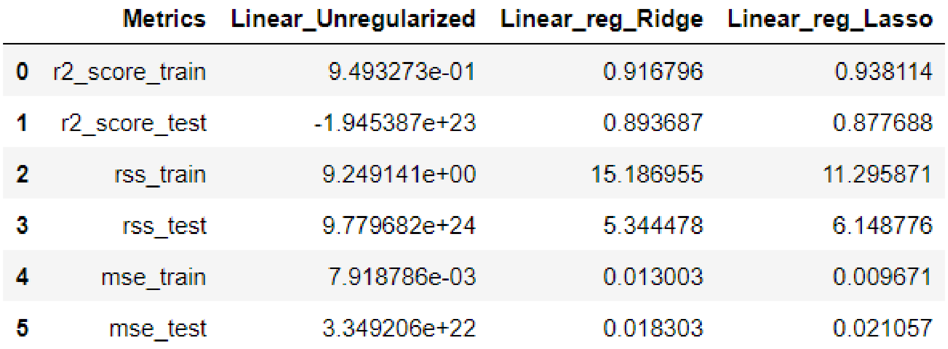
	Optimal value of alpha	Double of Optimal value of alpha
Ridge Regression	 Features ridge_coeff 2 OverallQual 0.297412 20 GrLivArea 0.241138 17 1stFlrSF 0.211746 3 OverallCond 0.190709 28 TotRmsAbvGrd 0.179420 18 2ndFlrSF 0.162554 23 FullBath 0.155383 32 GarageCars 0.147037 21 BsmtFullBath 0.121333 101 Neighborhood_StoneBr 0.120931	 Feature ridge_coeff 2 OverallQual 0.248857 20 GrLivArea 0.190164 28 TotRmsAbvGrd 0.168814 17 1stFlrSF 0.163394 3 OverallCond 0.150817 23 FullBath 0.148879 18 2ndFlrSF 0.135991 32 GarageCars 0.134366 33 GarageArea 0.113490 21 BsmtFullBath 0.103917
Lasso Regression	 Features lasso_coeff 20 GrLivArea 1.250513 141 RoofMatl_WdShngl 0.830497 135 RoofMatl_CompShg 0.758039 139 RoofMatl_Tar&Grv 0.707657 136 RoofMatl_Membran 0.686333 140 RoofMatl_WdShake 0.641092 138 RoofMatl_Roll 0.599780 137 RoofMatl_Metal 0.573273 2 OverallQual 0.380803 60 MSZoning_FV 0.375005	 Feature lasso_coeff 20 GrLivArea 0.692861 2 OverallQual 0.538423 32 GarageCars 0.226418 28 TotRmsAbvGrd 0.137555 9 BsmtQual 0.132846 3 OverallCond 0.121179 30 FireplaceQu 0.113192 27 KitchenQual 0.097741 21 BsmtFullBath 0.087393 23 FullBath 0.085197

Observations: As expected, the R2 score values got dropped in both the models. Co-efficient got reduced and in both the models there is change in order of the features effecting the SalePrice.

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:



	Metrics	Linear_Unregularized	Linear_reg_Ridge	Linear_reg_Lasso
0	r2_score_train	9.493273e-01	0.916796	0.938114
1	r2_score_test	-1.945387e+23	0.893687	0.877688
2	rss_train	9.249141e+00	15.186955	11.295871
3	rss_test	9.779682e+24	5.344478	6.148776
4	mse_train	7.918786e-03	0.013003	0.009671
5	mse_test	3.349206e+22	0.018303	0.021057

In Lasso regression, the r2 score of train is very good compared to Ridge regression, but the Ridge does better performance on test set.

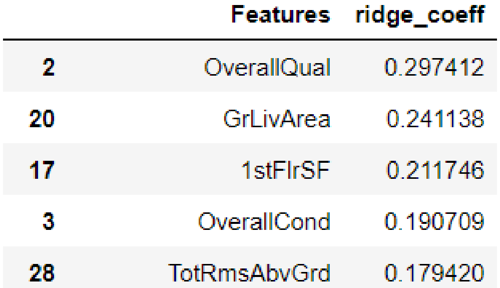
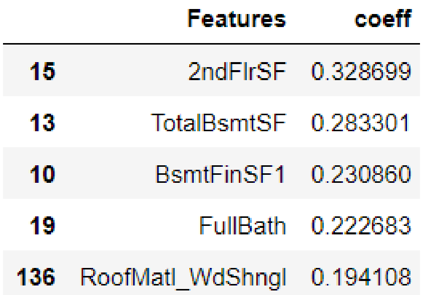
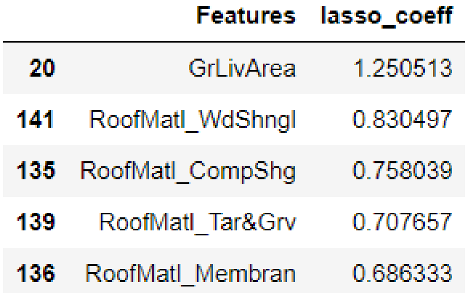
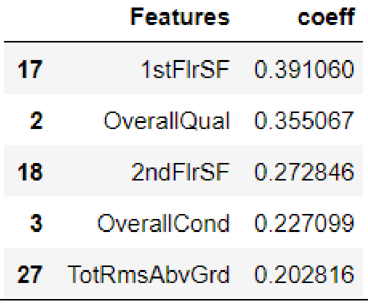
Lasso regression does the feature selection and help reduce the dimensionality which helps to interpret the model bit better compared to Ridge regression which tends to include all the available feature by keep their co-efficients close to zero but not zero.

Being said that, in real estate – especially housing domain, many factors or all factors contribute to the SalePrice but some have less effect compared to others. So, I would select Ridge regression.

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:

<table>
<tr><th></th><th>Actual - Features</th><th>Afte removing Top 5 features</th></tr>
<tr><td>Ridge Regression</td><td>
<table>
<tr><th></th><th>Features</th><th>ridge_coeff</th></tr>
<tr><td>2</td><td>OverallQual</td><td>0.297412</td></tr>
<tr><td>20</td><td>GrLivArea</td><td>0.241138</td></tr>
<tr><td>17</td><td>1stFlrSF</td><td>0.211746</td></tr>
<tr><td>3</td><td>OverallCond</td><td>0.190709</td></tr>
<tr><td>28</td><td>TotRmsAbvGrd</td><td>0.179420</td></tr>
</table></td><td>
<table>
<tr><th></th><th>Features</th><th>coeff</th></tr>
<tr><td>15</td><td>2ndFlrSF</td><td>0.328699</td></tr>
<tr><td>13</td><td>TotalBsmtSF</td><td>0.283301</td></tr>
<tr><td>10</td><td>BsmtFinSF1</td><td>0.230860</td></tr>
<tr><td>19</td><td>FullBath</td><td>0.222683</td></tr>
<tr><td>136</td><td>RoofMatl_WdShngl</td><td>0.194108</td></tr>
</table></td></tr>
<tr><td>Lasso Regression</td><td>
<table>
<tr><th></th><th>Features</th><th>lasso_coeff</th></tr>
<tr><td>20</td><td>GrLivArea</td><td>1.250513</td></tr>
<tr><td>141</td><td>RoofMatl_WdShngl</td><td>0.830497</td></tr>
<tr><td>135</td><td>RoofMatl_CompShg</td><td>0.758039</td></tr>
<tr><td>139</td><td>RoofMatl_Tar&Grv</td><td>0.707657</td></tr>
<tr><td>136</td><td>RoofMatl_Membran</td><td>0.686333</td></tr>
</table></td><td>
<table>
<tr><th></th><th>Features</th><th>coeff</th></tr>
<tr><td>17</td><td>1stFlrSF</td><td>0.391060</td></tr>
<tr><td>2</td><td>OverallQual</td><td>0.355067</td></tr>
<tr><td>18</td><td>2ndFlrSF</td><td>0.272846</td></tr>
<tr><td>3</td><td>OverallCond</td><td>0.227099</td></tr>
<tr><td>27</td><td>TotRmsAbvGrd</td><td>0.202816</td></tr>
</table></td></tr>
</table>

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?

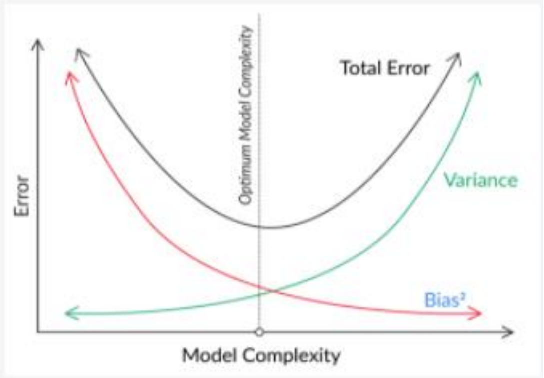
Answer:

Model is robust and generalizable when the variance and bias are rightly balanced. An extreme complex model will have high variance and it tends to memorize the noise in the data along with the important patterns.

This would usually result in an overfit model which means the performance is high on trained data, but this fails on unseen data.

And for a simple model, the bias will be too high, and the model fails to identify simple patterns in the data in simple word, underfitting.

So, to make the model robust and generalizable, it is important to find the correct balance between variance and bias. (Btw simple and complex model)



This can be achieved by incorporating regularization techniques.

Regularization helps to bring down the model complexity by shrinking the model coefficient towards 0. Regularization can be done by adding the penalty to the cost function.

Cost = RSS + Penalty

There are two commonly used techniques:

- Ridge Regression
- Lasso Regression

Ridge regression: This model adds the penalty with of summation of squares of co-efficients, multiplied with hyperparameter (lambda or alpha)

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j{}^2$$

Lasso Regression: In this model, the penalty is the absolute sum of all co-efficients multiplied with hyperparameter.

$$C = \sum_{1=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{F} |\, \beta_j|$$

Hyperparameter is used to fine tune the penalty so we can control the regularization of the model.

Higher value of lambda results in higher penalty and more regularization which results in simple model & underfitting mode. Whereas if lambda is too low, then model will be complex and over fitting.

So, by tuning the hyperparameter, we can bring the right balance on the complexity and result in robust and generalized model.

Implication on the model accuracy:
Accuracy alone on training data does not define the performance of the model. As we have seen in the housing case study, for an unregularized but, the r2 score value on training data is close to 0.95 but on the test data, it got negative r2 score.

A robust and generalized model should give more accurate performance on the unseen data, which in turn gives high accuracy and reliability.