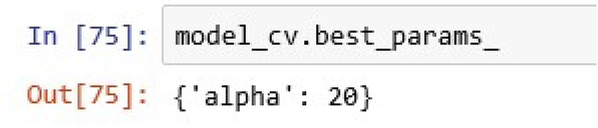
R.Santhosh_AdvancedRegressionAssignment_SubjectiveQuestions

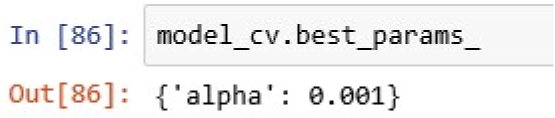
Question 1

What is the optimal value of alpha for ridge and lasso regression?

The optimal values of alpha for ridge and lasso regression are 20 and 0.001 respectively.



```
In [75]: model_cv.best_params_
Out[75]: {'alpha': 20}
```



```
In [86]: model_cv.best_params_
Out[86]: {'alpha': 0.001}
```

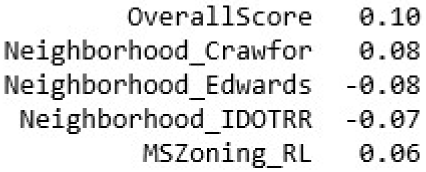
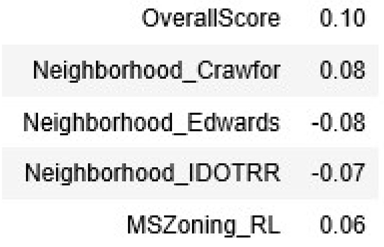
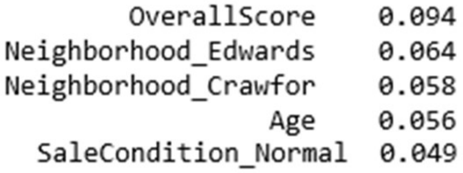
What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

	optimal values of alpha	**R^2(Train)**	**R^2(Test)**
Ridge	20	0.87	0.83
Lasso	0.001	0.86	0.83

	2 * optimal values of alpha	**R^2(Train)**	**R^2(Test)**
Ridge	40	0.86	0.83
Lasso	0.002	0.84	0.82

There is slight reduction of R^2 value if the value of alpha is doubled.

What will be the most important predictor variables after the change is implemented?

Top 5 Features with their Coefficients		
	Ridge	**Lasso**
optimal values of alpha	 OverallScore 0.10 Neighborhood_Crawfor 0.08 Neighborhood_Edwards -0.08 Neighborhood_IDOTRR -0.07 MSZoning_RL 0.06	 OverallScore 0.10 Neighborhood_Crawfor 0.08 Neighborhood_Edwards -0.08 Neighborhood_IDOTRR -0.07 MSZoning_RL 0.06
2 * optimal values of alpha	 OverallScore 0.094 Neighborhood_Edwards 0.064 Neighborhood_Crawfor 0.058 Age 0.056 SaleCondition_Normal 0.049	OverallScore 0.112 Neighborhood_Crawfor 0.071 Neighborhood_Edwards 0.069 Neighborhood_Somerst 0.068 MSZoning_RL 0.061

******* Here, OverallScore is a derived metric and it is given by OverallScore = OverallQual * OverallCond * GrLivArea

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?

When compared to Ridge, Lasso eliminates more number of features by penalizing the coefficients shrinking to zero. But Ridge regression make the coefficients tends to zero (but not zero), resulting in more number of features that makes model complex.

Thus Lasso which eliminates the features with better R-squared value will be chosen.

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?

Top 5 Predictor Variables using Lasso after Droping 5 Important Predictor Variables (with their coefficients) :

```
['RoofMatl_CompShg'    1.033]

['RoofStyle_Shed'          0.964]

['RoofStyle_Gable'          0.944]

['RoofStyle_Mansard'    0.933]

['RoofMatl_Membran'   0.923]]
```

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?

The robustness of the model can be ensured at different levels start from data preparation to model building. At the data level, proper measures for bias mitigation, data balancing, appropriate choice of features, using diverse training data can be done.

At the model building level, methods like cross-validation can generalize the model to unseen data.

A robust and generalized model may have less accuracy, in particular on training set. However, the robust model could have better accuracy on unseen data of test set.