Advanced Regression Assignment – Part 2 – Subjective questions

(Submission by Uday Kumar Adavi)

Question 1:

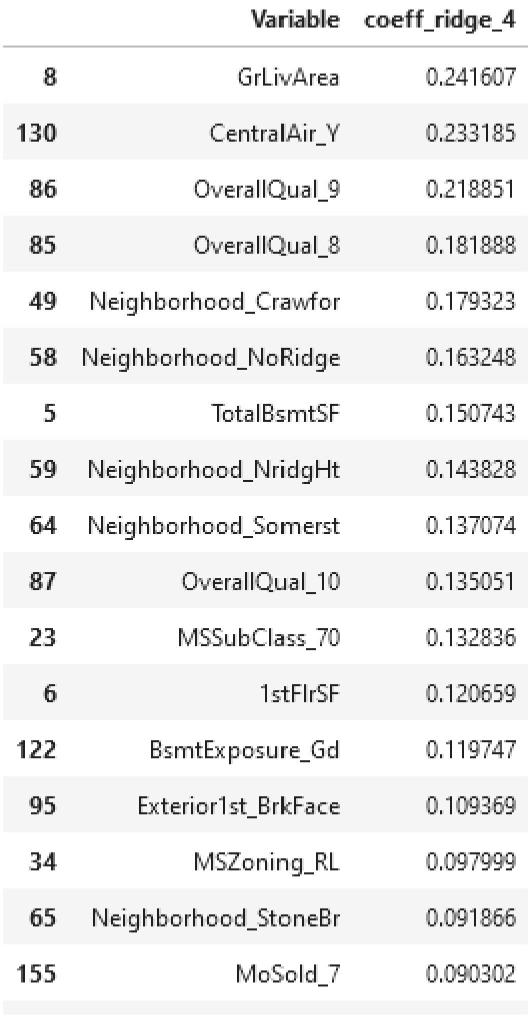
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Response to Question 1:

For the best Ridge regression model, the optimal value of alpha was 10. For the best Lasso regression model, the optimal value of alpha was 0.002. In principle, if we double the value of alpha for Ridge or Lasso, we will expect the regularisation penalty in the Cost function to increase. Hence, we will expect a less complex model and the values of the parameters moving closer to zero or zero in the case of Lasso. Also, the performance will deteriorate as we picked the optimal alpha for optimal performance.

For Ridge, R2 changed from 86.526% to 86.16% and RMSE from 0.37593 to 0.38091.

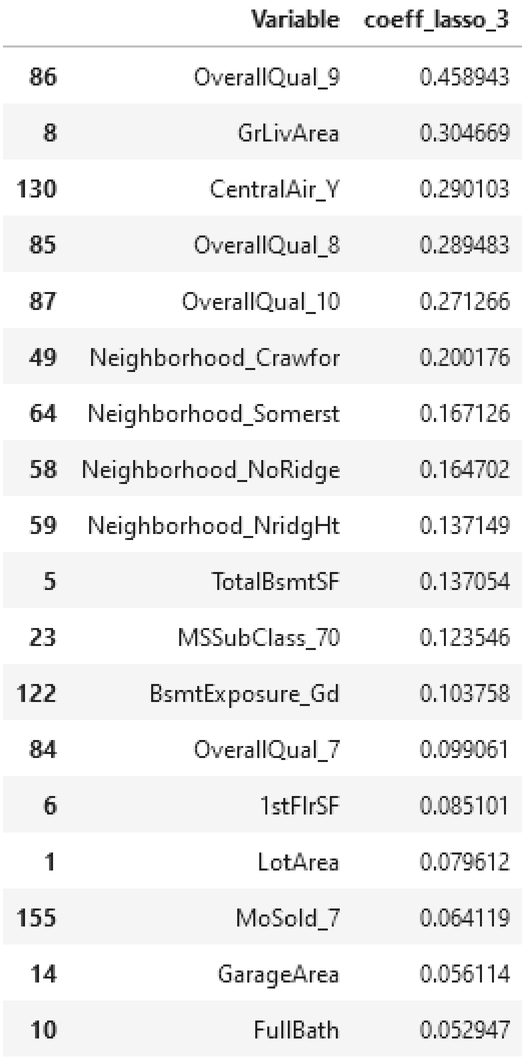
Top variables in Ridge model after change alpha to 2*optimal:



	Variable	coeff_ridge_4
8	GrLivArea	0.241607
130	CentralAir_Y	0.233185
86	OverallQual_9	0.218851
85	OverallQual_8	0.181888
49	Neighborhood_Crawfor	0.179323
58	Neighborhood_NoRidge	0.163248
5	TotalBsmtSF	0.150743
59	Neighborhood_NridgHt	0.143828
64	Neighborhood_Somerst	0.137074
87	OverallQual_10	0.135051
23	MSSubClass_70	0.132836
6	1stFlrSF	0.120659
122	BsmtExposure_Gd	0.119747
95	Exterior1st_BrkFace	0.109369
34	MSZoning_RL	0.097999
65	Neighborhood_StoneBr	0.091866
155	MoSold_7	0.090302

For Lasso, R2 changed from 86.357% to 85.257% and RMSE from 0.37827 to 0.3932.

Top variables in Lasso model after change alpha to 2*optimal:



	Variable	coeff_lasso_3
86	OverallQual_9	0.458943
8	GrLivArea	0.304669
130	CentralAir_Y	0.290103
85	OverallQual_8	0.289483
87	OverallQual_10	0.271266
49	Neighborhood_Crawfor	0.200176
64	Neighborhood_Somerst	0.167126
58	Neighborhood_NoRidge	0.164702
59	Neighborhood_NridgHt	0.137149
5	TotalBsmtSF	0.137054
23	MSSubClass_70	0.123546
122	BsmtExposure_Gd	0.103758
84	OverallQual_7	0.099061
6	1stFlrSF	0.085101
1	LotArea	0.079612
155	MoSold_7	0.064119
14	GarageArea	0.056114
10	FullBath	0.052947

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?

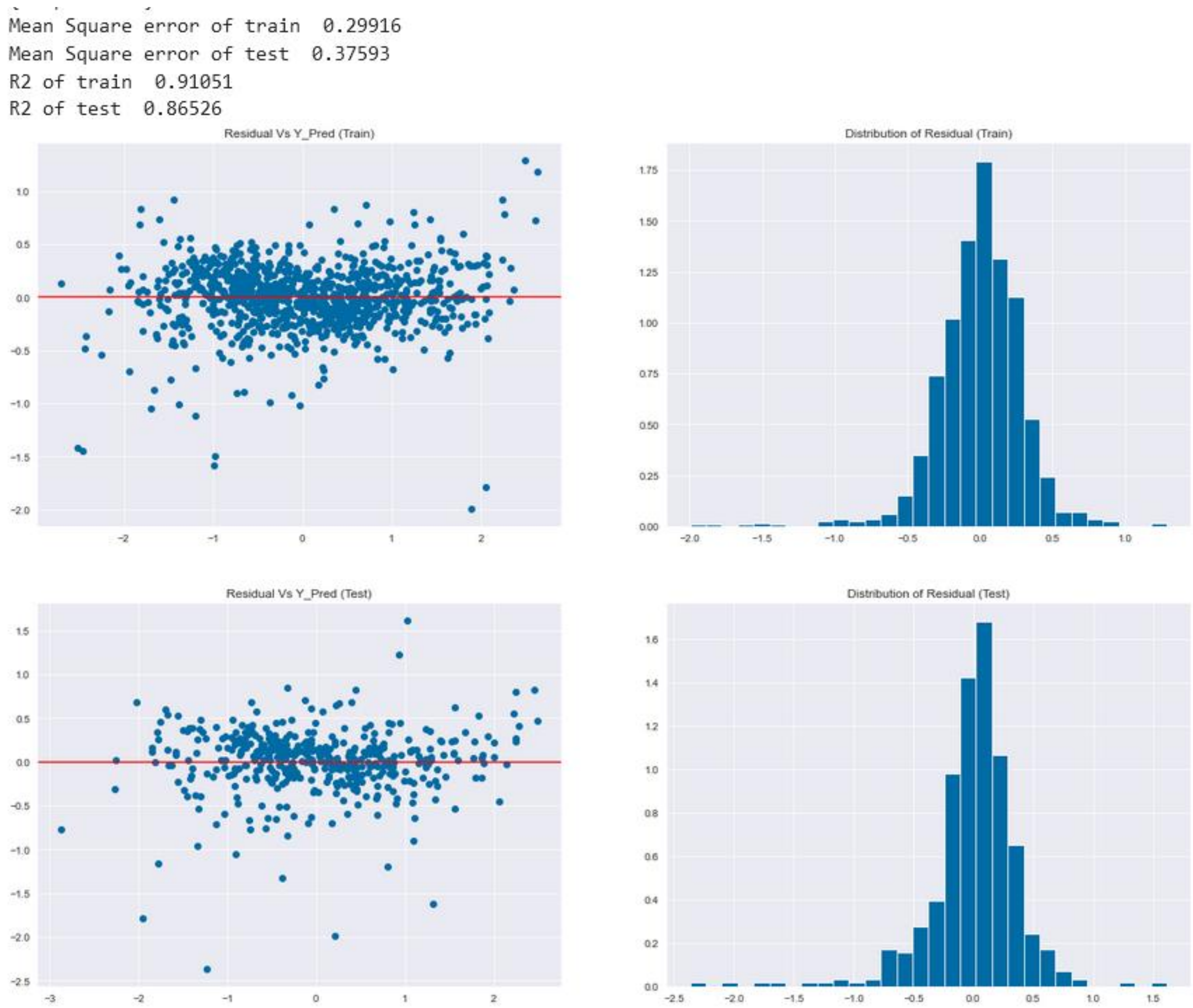
Response to Question 2:

To identify the final model, we need to look at:

- the model metric scores for test
    - Rsq
    - RMSE
- compare test and train values to check for overfit. we'd like to see test results that are closest to train for a more stable model that is not overfit
- model validation through error analysis. All models should comply with assumptions of linear regression

Keeping all these together, the final model we will select is the third model from Ridge regression. Has the better scores for test and the minimal difference from train metrics, while also fulfilling error analysis needs. (compared with the best Lasso regression model which had (on test) R2 of 86.35% and RMSE of 0.37827)

Performance & validation of the final model picked(3rd model from Ridge regression):



```
Mean Square error of train  0.29916
Mean Square error of test  0.37593
R2 of train  0.91051
R2 of test  0.86526
```

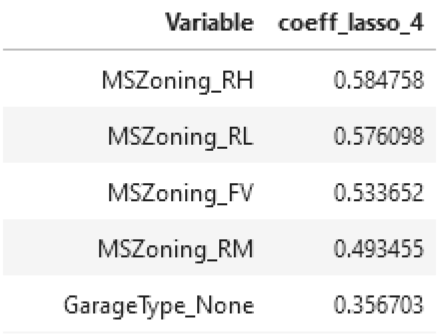
Question 3:

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?

Response to Question 3:

In the best Lasso model, the top 5 variables are: ‘OverallQual_9’, ‘OverallQual_10’, ‘OverallQual_8’, ‘CentralAir_Y’, ‘GrLivArea’.

After removing these variables and re-training the model including grid search to find the optimal alpha, the (on test) R2 is 86.5% and RMSE is 0.37636. The optimal value of alpha is 0.0006. The top 5 variables now are:



Variable	coeff_lasso_4
MSZoning_RH	0.584758
MSZoning_RL	0.576098
MSZoning_FV	0.533652
MSZoning_RM	0.493455
GarageType_None	0.356703

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?

Response to Question 4:

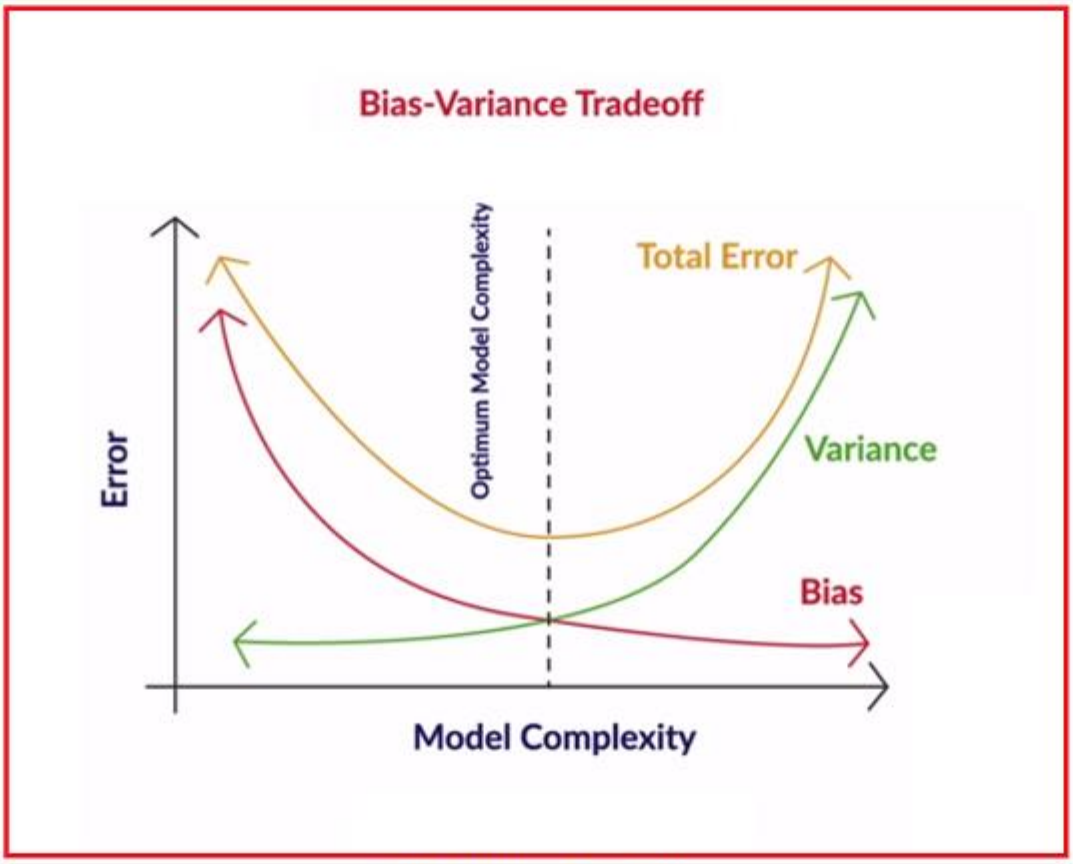
Accuracy Vs generalisability is the constant trade off we face with in model selection. Robustness and generalisability needs us to be more consistent in the results as we change the input data. This is the ‘variance’. For robustness and generalisability, we need lower variance. This can be achieved by keeping the model simple and less complex. Simpler models have less variance, are more stable and less sensitive to input data. But simpler models perform poorly from accuracy perspective in the real world and tend to have high bias. This is where we need to pick the right optimal point where we have the lowest variance for the lowest bias possible.

From a practical perspective, we look at the key metrics across test and train and ensure there is minimal difference between train and test. Cross validation also helps reduce overfitting.

When we have high bias issues, we need more data/input points, generally, there aren’t enough patterns in the data that are helping explain the variance if high bias is an issue. Adding polynomial features can help improve performance and so can some transformations. If we are using regularisation, then we can reduce the weightage.

When we have high variance, we should be looking at regularisation techniques, use robust variable selection, get more training data

Illustration from from UpGrad lecture/notes, demonstrated the bias variance trade-off:



Bias Variance Tradeoff