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?

Ans:
The optimal value of alpha for

Ridge	2
Lasso	0.0002

If we double the value of Ridge and lasso, Below is the changes in co-efficients and overall model performance.

Overall Performance of Model

Ridge – 2
Lasso – 0.0002

	Metric	Lasso Alpha = 0.0002	Ridge Alpha = 2
0	R2 Score (Train)	0.9110	0.9109
1	R2 Score (Test)	0.8783	0.8787
2	RSS (Train)	8.2117	8.2186
3	RSS (Test)	4.9864	4.9686
4	MSE (Train)	0.0993	0.0993
5	MSE (Test)	0.1180	0.1178

Ridge – 4
Lasso – 0.0004

	Metric	Lasso Alpha = 0.0004	Ridge Alpha = 4
0	R2 Score (Train)	0.9103	0.9104
1	R2 Score (Test)	0.8809	0.8807
2	RSS (Train)	8.2703	8.2665
3	RSS (Test)	4.8808	4.8864
4	MSE (Train)	0.0996	0.0996
5	MSE (Test)	0.1168	0.1168

Ridge Co-efficient Values

Ridge – 2

	Features	Coefficient
12	MSZoning_RL	0.1768
13	MSZoning_RM	0.1396
10	MSZoning_FV	0.1099
6	2ndFlrSF	0.0888
0	OverallQual	0.0714
5	1stFlrSF	0.0712
4	TotalBsmtSF	0.0475
1	OverallCond	0.0457
3	BsmtFinSF1	0.0363
11	MSZoning_RH	0.0361
2	YearBuilt	0.0348
8	GarageCars	0.0332
7	GrLivArea	0.0281
14	BldgType_Duplex	-0.0222
9	DiffYearBuilt	-0.0348

Ridge – 4

	Features	Coefficient
12	MSZoning_RL	0.1591
13	MSZoning_RM	0.1242
10	MSZoning_FV	0.1001
0	OverallQual	0.0717
6	2ndFlrSF	0.0702
5	1stFlrSF	0.0579
4	TotalBsmtSF	0.0475
1	OverallCond	0.0460
7	GrLivArea	0.0460
3	BsmtFinSF1	0.0363
2	YearBuilt	0.0349
8	GarageCars	0.0329
11	MSZoning_RH	0.0321
14	BldgType_Duplex	-0.0221
9	DiffYearBuilt	-0.0349

Lasso Co-efficient Values

Lasso – 0.0002

	Features	Coefficient
11	MSZoning_RL	0.186151
12	MSZoning_RM	0.147299
6	2ndFlrSF	0.116587
9	MSZoning_FV	0.114955
5	1stFlrSF	0.092388
0	OverallQual	0.071743
2	YearBuilt	0.067347
4	TotalBsmtSF	0.050304
1	OverallCond	0.045323
10	MSZoning_RH	0.038211
7	GarageCars	0.033279
3	BsmtFinSF1	0.032050
14	Exterior1st_VinylSd	0.028542
8	DiffYearBuilt	-0.000487
13	BldgType_Duplex	-0.021891

Lasso – 0.0004

	Features	Coefficient
12	MSZoning_RL	0.172753
13	MSZoning_RM	0.135152
10	MSZoning_FV	0.107325
6	2ndFlrSF	0.087587
0	OverallQual	0.072461
5	1stFlrSF	0.071718
2	YearBuilt	0.066275
4	TotalBsmtSF	0.049502
1	OverallCond	0.045464
11	MSZoning_RH	0.035256
8	GarageCars	0.033119
3	BsmtFinSF1	0.032254
7	GrLivArea	0.027829
9	DiffYearBuilt	-0.001018
14	BldgType_Duplex	-0.021457

As we able to see overall model performance there is not much change. But we observe very small change in value of co-efficients in ridge and lasso.

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?

Ans: Ridge and lasso co-efficient are dependent on lambda value. In the process of regularization that is increasing the value of lambda will point co-efficient to 0 in lasso and in ridge the value close to 0.

	Metric	Lasso Alpha = 0.0002	Ridge Alpha = 2
0	R2 Score (Train)	0.9110	0.9109
1	R2 Score (Test)	0.8783	0.8787
2	RSS (Train)	8.2117	8.2186
3	RSS (Test)	4.9864	4.9686
4	MSE (Train)	0.0993	0.0993
5	MSE (Test)	0.1180	0.1178

In our dataset as you can see lasso and ridge have approximately close MSE and RSS. Since lasso helps in feature reduction (as the coefficient value of one of the feature become 0).

Lasso has a better edge over ridge if number of features in data are high.

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?

Ans: top five lasso co-efficient are related to Zoning classification feature(Residential Low Density, Medium Density & Floating Village Residentials) and first and second floor square feet

	Features	rfe_support	rfe_ranking	Coefficient
11	MSZoning_RL	True	1	0.186151
12	MSZoning_RM	True	1	0.147299
6	2ndFlrSF	True	1	0.116587
9	MSZoning_FV	True	1	0.114955
5	1stFlrSF	True	1	0.092388

If these features are not available then we can use the remaining features as shown below.

That is next top 5 predictors are Overall material and finish of the house, age of house, Total square feet of basement area, overall condition of the house and Zoning Classification Feature with Residential High Density.

0	OverallQual	True	1	0.071743
2	YearBuilt	True	1	0.067347
4	TotalBsmtSF	True	1	0.050304
1	OverallCond	True	1	0.045323
10	MSZoning_RH	True	1	0.038211
7	GarageCars	True	1	0.033279
3	BsmtFinSF1	True	1	0.032050
14	Exterior1st_VinylSd	True	1	0.028542
8	DiffYearBuilt	True	1	-0.000487
13	BldgType_Duplex	True	1	-0.021891

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?

Ans:

Occam's Razor gives 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 the following reasons:-

1) 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
2) Simpler models are usually more generic and more widely applicable.
3) Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
4) Simpler models are more robust compared to complex models.
5) Complex models tend to change wildly with changes in the training data set
6) Simple models have low variance, high bias and complex models have low bias, high variance

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.

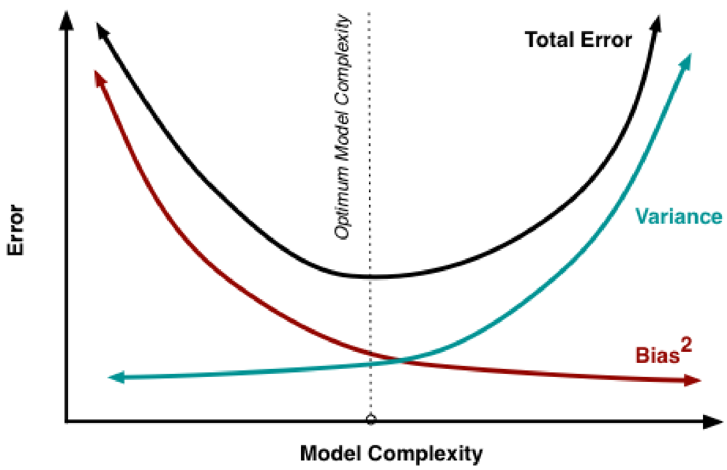
Also, Making a model simple leads to Bias-Variance Trade-off:

- 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 the 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 the 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