Assignment-based Subjective Questions

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 values of alpha for Ridge regression is 20 and R2 is 0.926

   The optimal values of alpha for Lasso regression is 0.001 and R2 is 0.925

   After doubling the values of alpha of the models are below

   The optimal values of alpha for Ridge regression is 40 and R2 is 0.925

   The optimal values of alpha for Lasso regression is 0.002 and R2 is 0.923

   There is small change in R2 values after doubling the alpha values for both Ridge and Lasso models. The new model is created and demonstrated in the Jupiter notebook. Below are the changes in the co-efficient after doubling the alpha values.

Ridge Coefficient (Initial alpha)

	Features	Coefficient
0	MSSubClass	-0.0252
1	LotArea	0.0189
2	OverallQual	0.0787
3	OverallCond	0.0434
4	BsmtFinSF1	0.0246
5	BsmtUnfSF	-0.0060
6	TotalBsmtSF	0.0541
7	1stFlrSF	0.0338
8	2ndFlrSF	0.0322
9	GrLivArea	0.0634
10	BsmtFullBath	0.0131
11	FullBath	0.0161
12	HalfBath	0.0157
13	GarageCars	0.0339
14	WoodDeckSF	0.0128
15	IsRemodelled	-0.0163
16	BuiltOrRemodelAge	-0.0237
17	OldOrNewGarage	0.0122
18	d_BsmtQual	0.0152
19	d_BsmtExposure	0.0124
20	d_HeatingQC	0.0120
21	d_KitchenQual	0.0137
22	d_FireplaceQu	0.0194
23	d_GarageFinish	0.0107
24	d_HouseStyle	0.0146
25	d_SaleCondition	0.0166
26	MSZoning_FV	0.0414
27	MSZoning_RH	0.0203

Ridge Coefficient (After doubling alpha)

	Features	Coefficient
0	MSSubClass	-0.0243
1	LotArea	0.0186
2	OverallQual	0.0770
3	OverallCond	0.0428
4	BsmtFinSF1	0.0262
5	BsmtUnfSF	-0.0037
6	TotalBsmtSF	0.0504
7	1stFlrSF	0.0361
8	2ndFlrSF	0.0328
9	GrLivArea	0.0607
10	BsmtFullBath	0.0134
11	FullBath	0.0172
12	HalfBath	0.0163
13	GarageCars	0.0329
14	WoodDeckSF	0.0127
15	IsRemodelled	-0.0163
16	BuiltOrRemodelAge	-0.0238
17	OldOrNewGarage	0.0120
18	d_BsmtQual	0.0157
19	d_BsmtExposure	0.0125
20	d_HeatingQC	0.0121
21	d_KitchenQual	0.0151
22	d_FireplaceQu	0.0198
23	d_GarageFinish	0.0110
24	d_HouseStyle	0.0141
25	d_SaleCondition	0.0171
26	MSZoning_FV	0.0328
27	MSZoning_RH	0.0156

Lasso Coefficient (Initial alpha)

	Features	Coefficient
0	MSSubClass	-0.0232
1	LotArea	0.0173
2	OverallQual	0.0837
3	OverallCond	0.0440
4	BsmtFinSF1	0.0264
5	BsmtUnfSF	-0.0038
6	TotalBsmtSF	0.0541
7	1stFlrSF	0.0043
8	GrLivArea	0.1022
9	BsmtFullBath	0.0127
10	FullBath	0.0132
11	HalfBath	0.0138
12	GarageCars	0.0347
13	WoodDeckSF	0.0123
14	IsRemodelled	-0.0161
15	BuiltOrRemodelAge	-0.0240
16	OldOrNewGarage	0.0108
17	d_BsmtQual	0.0120
18	d_BsmtExposure	0.0125
19	d_HeatingQC	0.0107
20	d_KitchenQual	0.0132
21	d_FireplaceQu	0.0186
22	d_GarageFinish	0.0107
23	d_HouseStyle	0.0119
24	d_SaleCondition	0.0168
25	MSZoning_FV	0.0364
26	MSZoning_RH	0.0157
27	MSZoning_RL	0.0491

Lasso Coefficient (After doubling alpha)

	Features	Coefficient
0	MSSubClass	-0.0205
1	LotArea	0.0164
2	OverallQual	0.0857
3	OverallCond	0.0431
4	BsmtFinSF1	0.0297
5	TotalBsmtSF	0.0501
6	1stFlrSF	0.0040
7	GrLivArea	0.1026
8	BsmtFullBath	0.0126
9	FullBath	0.0121
10	HalfBath	0.0129
11	GarageCars	0.0358
12	WoodDeckSF	0.0119
13	IsRemodelled	-0.0157
14	BuiltOrRemodelAge	-0.0243
15	OldOrNewGarage	0.0094
16	d_BsmtQual	0.0098
17	d_BsmtExposure	0.0121
18	d_HeatingQC	0.0101
19	d_KitchenQual	0.0139
20	d_FireplaceQu	0.0182
21	d_GarageFinish	0.0120
22	d_HouseStyle	0.0092
23	d_SaleCondition	0.0170
24	MSZoning_FV	0.0289
25	MSZoning_RH	0.0104
26	MSZoning_RL	0.0376
27	Neighborhood_BrkSide	0.0127

Most important predictor variables are below after the change is implemented

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. MSZoning_RL
5. OverallCond

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:

Optimum Lambda values

- The optimal values of alpha for Ridge regression is 20
- The optimal values of alpha for Lasso regression is 0.001

Mean squared error values

- Mean square error value for Ridge regression is 0.1266
- Mean square error value for Lasso regression is 0.1267

The mean square error values for both Ridge and Lasso regression are almost same. So, in terms of accuracy, both Ridge and Lasso regression will give result with almost same accuracy.

However, Lasso regression will be chosen as better option in this scenario as it reduces features by making some of the coefficient values of features to zero. This makes interpretation of analysis result easier.

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 models are given below

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. MSZoning_RL
5. OverallCond

New model is built after removing the above 5 features. After removing the above 5 features, R2 for the model is reduced to 0.90 and mean squared error is increased to 0.14

The below features are new top 5 predictors.

Features	rfe_support	rfe_ranking	Coefficient
2ndFlrSF	True	1	0.115116
1stFlrSF	True	1	0.105435
BsmtFinSF1	True	1	0.052809
GarageCars	True	1	0.040950
d_KitchenQual	True	1	0.038222

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

Answer:

The model should be simple as necessary. As per Occam's razor, if two models give similar performance in the training and test data, pick the model which is simpler with less number of feature and less complexity.

The simpler model is usually more generic and robust. They are widely applicable.

The simpler models require fewer training data for effective training than the more complex one and hence simpler models are easier to train. Simpler models have low variance and high bias.

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. Complex model should

be avoided as any change in training data will have significant change in the model. Complex modes have high variance and low bias. Complex models lead to over-fitting. So it is not robust and generalizable.

There should be Bias-Variance trade-off while making the model simpler. The implication of the very simpler model is that it will reduce accuracy in the result as it will leads to under-fitting. So, balancing is required between simpler and complex model.

Regularization can be used to make the model simpler without much compromise on bias. It helps to strike balance between variance and bias. Regularization involves addition of regularization term to the cost that adds up the absolute values or square of the parameters of the model.

Bias quantifies how accurately the model can describe the actual task. 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.