Problem Statement - Part II

Assignment Part-II

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?

- alpha for ridge = 100 & alpha for Lasso = 0.01
- There is an increase in the mean squared error for Ridge & no change in mean squared error for Lasso when we double the value of alpha.
- There is no change in top 10 predictor variables even we double the value of alpha for both Ridge & Lasso.

Ridge:

Alpha = 100, mean_squared_error = 474135173.12

	Features	rfe_support	rfe_ranking	Coefficient
3	GrLivArea	True	1	12136.3129
0	OverallQual	True	1	11337.3660
10	Neighborhood_NridgHt	True	1	7401.7513
4	GarageCars	True	1	7345.6921
1	BsmtFinSF1	True	1	7122.7039
2	TotalBsmtSF	True	1	5743.3739
5	d_ExterQual	True	1	5671.0762
6	d_BsmtExposure	True	1	5593.1929
7	d_KitchenQual	True	1	5288.9946
9	Neighborhood_Crawfor	True	1	4767.0637
11	Exterior1st_Stucco	True	1	2966.7499
14	Exterior2nd_VinylSd	True	1	2355.0232
12	Exterior1st_VinylSd	True	1	178.0426
13	Exterior2nd_Stucco	True	1	-2982.2036
8	d_BldgType	True	1	-5973.2815

Alpha = 200, mean_squared_error = 476458717.05

	Features	rfe_support	rfe_ranking	Coefficient
3	GrLivArea	True	1	12136.3129
0	OverallQual	True	1	11337.3660
10	Neighborhood_NridgHt	True	1	7401.7513
4	GarageCars	True	1	7345.6921
1	BsmtFinSF1	True	1	7122.7039
2	TotalBsmtSF	True	1	5743.3739
5	d_ExterQual	True	1	5671.0762
6	d_BsmtExposure	True	1	5593.1929
7	d_KitchenQual	True	1	5288.9946
9	Neighborhood_Crawfor	True	1	4767.0637
11	Exterior1st_Stucco	True	1	2966.7499
14	Exterior2nd_VinylSd	True	1	2355.0232
12	Exterior1st_VinylSd	True	1	178.0426
13	Exterior2nd_Stucco	True	1	-2982.2036
8	d_BldgType	True	1	-5973.2815

Lasso

Alpha = 0.01, mean_squared_error = 506438455.12

	Features	rfe_support	rfe_ranking	Coefficient
3	GrLivArea	True	1	29782.924485
0	OverallQual	True	1	11889.976153
14	Exterior2nd_VinylSd	True	1	8844.630001
10	Neighborhood_NridgHt	True	1	8633.454776
4	GarageCars	True	1	8493.097584
1	BsmtFinSF1	True	1	7542.542267
6	d_BsmtExposure	True	1	6426.906193
2	TotalBsmtSF	True	1	6175.247235
9	Neighborhood_Crawfor	True	1	5443.713383
5	d_ExterQual	True	1	4907.693112
7	d_KitchenQual	True	1	4581.008377
11	Exterior1st_Stucco	True	1	4492.015649
13	Exterior2nd_Stucco	True	1	-4560.254759
12	Exterior1st_VinylSd	True	1	-6296.384136
8	d_BldgType	True	1	-7807.588062

Alpha = 0.02, mean_squared_error = 506435126.20

	Features	rfe_support	rfe_ranking	Coefficient
3	GrLivArea	True	1	29780.138998
0	OverallQual	True	1	11890.031026
14	Exterior2nd_VinylSd	True	1	8844.062333
10	Neighborhood_NridgHt	True	1	8633.416073
4	GarageCars	True	1	8493.062570
1	BsmtFinSF1	True	1	7542.539187
6	d_BsmtExposure	True	1	6426.882197
2	TotalBsmtSF	True	1	6175.255915
9	Neighborhood_Crawfor	True	1	5443.679496
5	d_ExterQual	True	1	4907.720820
7	d_KitchenQual	True	1	4581.030760
11	Exterior1st_Stucco	True	1	4492.091486
13	Exterior2nd_Stucco	True	1	-4560.338277
12	Exterior1st_VinylSd	True	1	-6295.821439
8	d_BldgType	True	1	-7807.560680

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?

- The Mean Squared Error of Ridge is slightly lower than that of Lasso
- Optimal alpha value is higher for Ridge than Lasso
- Hence, we will be choosing Ridge

Question 3

After building the model, you realized 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?

MSSubclass,1stFlrSF,2ndFlrSF,Fireplaces,d_ExterQual

	Features	rfe_support	rfe_ranking	Coefficient
2	2ndFlrSF	True	1	21960.274097
1	1stFlrSF	True	1	21471.364174
9	Neighborhood_NridgHt	True	1	10574.893492
14	Exterior2nd_VinylSd	True	1	10367.964412
13	Exterior2nd_CmentBd	True	1	9286.171223
6	d_BsmtExposure	True	1	7799.727177
5	d_BsmtQual	True	1	7268.012251
4	d_ExterQual	True	1	5982.918872
8	d_KitchenQual	True	1	5966.537952
3	Fireplaces	True	1	5091.209158
7	d_BsmtFinType1	True	1	3563.055677
0	MSSubClass	True	1	-4049.510960
12	Exterior1st_Wd Sdng	True	1	-4292.792897
10	Exterior1st_CemntBd	True	1	-5681.620401
11	Exterior1st_VinylSd	True	1	-8534.474196

Question 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?

- The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable.
- Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.