Problem Statement - 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?

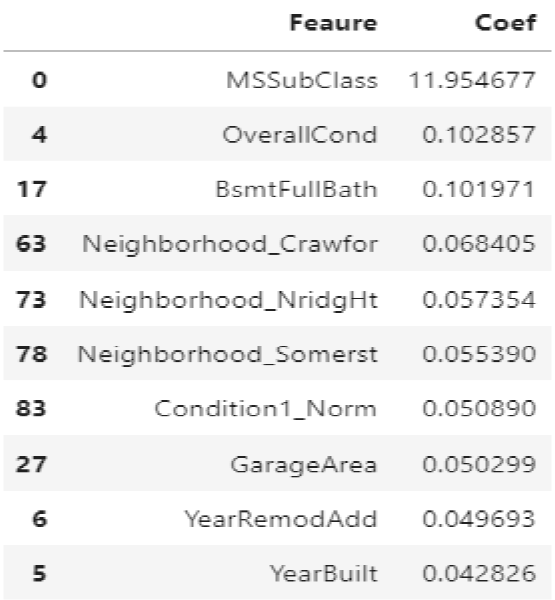
Answer:-

Most efficient alpha value for Lasso: {'alpha': 0.001}

Most efficient value for Ridge: {'alpha': 0.1}

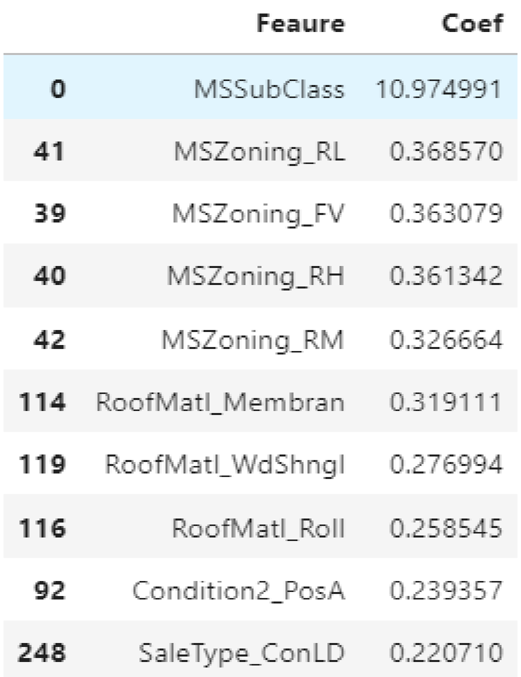
Features if we double the values of ridge and lasso coefficient:-

Lasso = 0.002



	Feaure	Coef
0	MSSubClass	11.954677
4	OverallCond	0.102857
17	BsmtFullBath	0.101971
63	Neighborhood_Crawfor	0.068405
73	Neighborhood_NridgHt	0.057354
78	Neighborhood_Somerst	0.055390
83	Condition1_Norm	0.050890
27	GarageArea	0.050299
6	YearRemodAdd	0.049693
5	YearBuilt	0.042826

For ridge = 0.2



	Feaure	Coef
0	MSSubClass	10.974991
41	MSZoning_RL	0.368570
39	MSZoning_FV	0.363079
40	MSZoning_RH	0.361342
42	MSZoning_RM	0.326664
114	RoofMatl_Membran	0.319111
119	RoofMatl_WdShngl	0.276994
116	RoofMatl_Roll	0.258545
92	Condition2_PosA	0.239357
248	SaleType_ConLD	0.220710

Changes after we double values of both ridge and lasso coefficients are:-

Lasso

R2 train 0.9174442574972889 → 0.8921940497658922

R2 test 0.855123192817482 → 0.8774018105216189

Ridge

R2 train 0.9575101480652919 → 0.9553979560480054

R2 test 0.749177935388393 → 0.7864035790812623

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?

Answer:-

So in our case ridge has better r2 for the training set and lasso has a better r2 for the test set. It will depend upon our use case.

If we require feature selection we will use lasso and we if we need an optimal value for the regression coefficient we will go for Ridge regression.

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?

Answer:-

1. MSSubClass_70
2. OverallQual
3. OverallCond
4. GarageArea
5. Neighborhood_Somerst

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?

Answer:-

- **The model should be as simple as possible because its accuracy can drop a few percentages but the model itself will remain robust and will not affect its performance too much.**

- **We just have to make sure that the model is not overfitting. Overfitting can cause high variance which will result in lower accuracy as even a little change in data will change the model prediction.**
- **So basically we have to find a fine balance between model accuracy and the simplicity of the model.**
- **We can apply various regularization techniques like lasso and ridge regression to find the balance between complexity and model accuracy.**