Assignment based Subjective Questions

Question 1

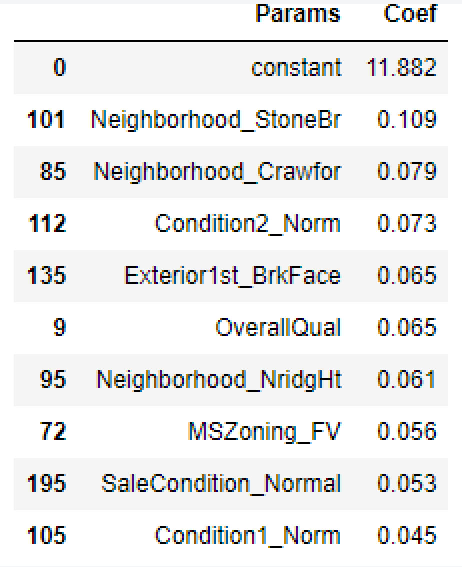
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

- Optimal values of alpha for
  - **Ridge:** 8.0
  - **Lasso:** 0.001

Ridge:

- **With Optimal Alpha value:**
  - The most predictor variables and its coefficients are as follows:

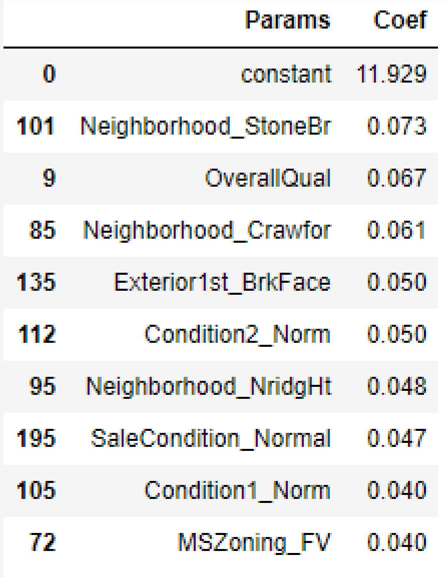


	Params	Coef
0	constant	11.882
101	Neighborhood_StoneBr	0.109
85	Neighborhood_Crawfor	0.079
112	Condition2_Norm	0.073
135	Exterior1st_BrkFace	0.065
9	OverallQual	0.065
95	Neighborhood_NridgHt	0.061
72	MSZoning_FV	0.056
195	SaleCondition_Normal	0.053
105	Condition1_Norm	0.045

- Scores are as follows:

SCORE TYPE	TRAIN	TEST
R2	0.9299	0.9143
RSS	7.4224	5.0136
MSE	0.0957	0.1200

- **With doubled Alpha:**
  - The most predictor variables and its coefficients are as follows:



	Params	Coef
0	constant	11.929
101	Neighborhood_StoneBr	0.073
9	OverallQual	0.067
85	Neighborhood_Crawfor	0.061
135	Exterior1st_BrkFace	0.050
112	Condition2_Norm	0.050
95	Neighborhood_NridgHt	0.048
195	SaleCondition_Normal	0.047
105	Condition1_Norm	0.040
72	MSZoning_FV	0.040

- Scores are as follows:

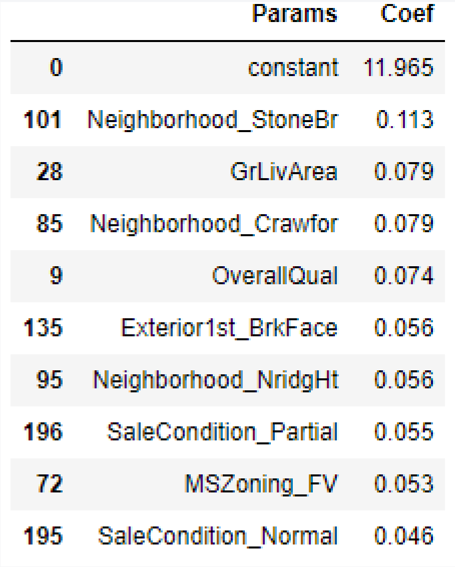
SCORE TYPE	TRAIN	TEST
R2	0.9248	0.9142
RSS	7.9627	5.0193
MSE	0.0098	0.0144

Conclusion:

1. We can observe that as the alpha got doubled, the top most predictors got changed (in the view of order) and their coefficients are tending towards closer to zero
2. We can see the slight reduction in the R2 score which means that model is losing its efficiency in predicting the model
3. RSS got increased in the training data set which means that discrepancy between the data and an estimation model got increased. But RSS of test data set is still performs well
4. MSE of doubled the optimal alpha got decreased alot.
5. As alpha increases the train and test scores reduces and the model efficiency gets reduced

Lasso:

- **With optimal alpha value:**
  - Top most predictors

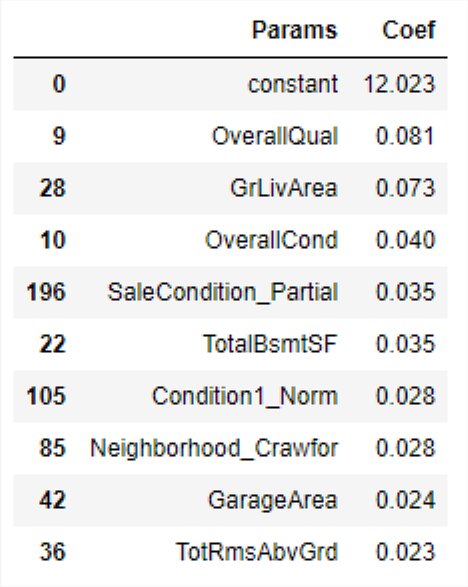


	Params	Coef
0	constant	11.965
101	Neighborhood_StoneBr	0.113
28	GrLivArea	0.079
85	Neighborhood_Crawfor	0.079
9	OverallQual	0.074
135	Exterior1st_BrkFace	0.056
95	Neighborhood_NridgHt	0.056
196	SaleCondition_Partial	0.055
72	MSZoning_FV	0.053
195	SaleCondition_Normal	0.046

- Scores are as follows:

SCORE TYPE	TRAIN	TEST
R2	0.9227	0.9197
RSS	8.1776	4.6968
MSE	0.1004	0.1161

- **With doubled alpha value:**
  - Top most predictors



	Params	Coef
0	constant	12.023
9	OverallQual	0.081
28	GrLivArea	0.073
10	OverallCond	0.040
196	SaleCondition_Partial	0.035
22	TotalBsmtSF	0.035
105	Condition1_Norm	0.028
85	Neighborhood_Crawfor	0.028
42	GarageArea	0.024
36	TotRmsAbvGrd	0.023

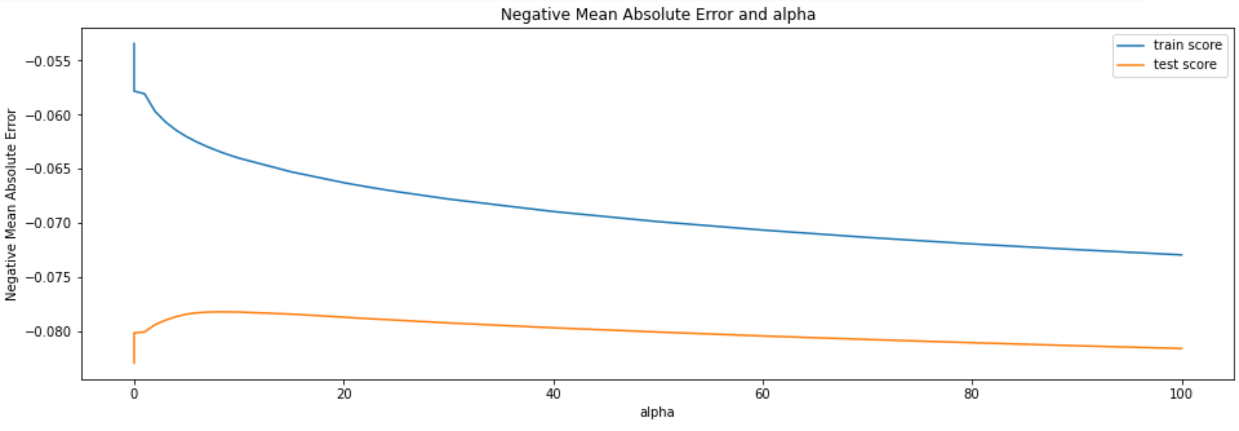
- Scores are as follows

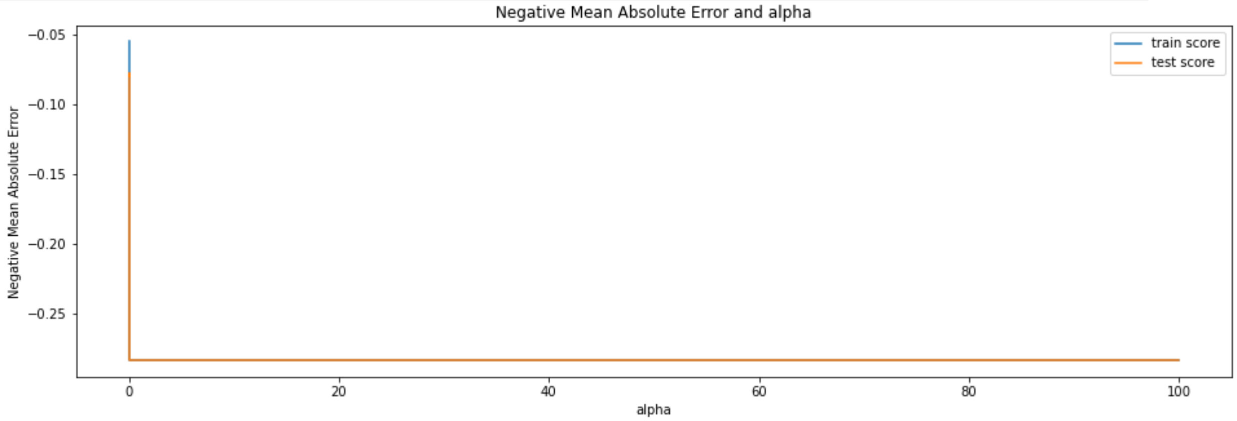
SCORE TYPE	TRAIN	TEST
R2	0.9032	0.9134
RSS	10.2434	5.0693
MSE	0.0126	0.0145

Conclusions:

1. R2 Score was significantly reduced when alpha got doubled which means that the model prediction and accuracy got reduced
2. RSS got increased which means that discrepancy between the data and an estimation model got increased.
3. MSE of doubled the optimal alpha got decreased alot.
4. Some of the top most predictors in optimal alpha are not available in the alpha when it got doubled and values are also tending to zero as alpha got doubled

- **Following graph shows how the training and test scores varies for different alphas of ridge and lasso models respectively**





We can observe that test and train scores reduce as alpha increases.

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:

- Although both the models perform better and provide best scores, I choose to apply lasso regression models.
- Because
  - As I observed, lasso performed slightly well on test data set compared to ridge
  - Lasso performs feature elimination as well by making the coefficients of predictors 0(Zero) which are not good for the model.
- Feature elimination is also a key factor for the model such that it removes the outliers and the noise in data.
- Hence, Lasso is a better option than the ridge regression model.

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?

Answer:

Earlier Model:

- Top most five predictors of lasso model earlier are:
    1. Neighborhood_StoneBr → Physical locations within Ames city limits are Stone Brook
    2. GrLivArea → Above grade (ground) living area square feet
    3. Neighbourhood_Crawfor → Physical locations within Ames city limits are Crawford
    4. OverallQual → Overall rating of material and finish of house
    5. Exterior1st_BrkFace → Exterior covering on house with Brick Face

After dropping the earlier five predictors of the lasso model. Current top five predictors of lasso model are:

1. MSZoning_FV → Floating Village Residential Zoning sale
2. MSZoning_RL → Residential Low Density Zoning sale
3. MSZoning_RH → Residential High Density Zoning sale
4. MSZoning_RM → Residential Medium Density Zoning sale
5. Condition2_PosA → Proximity to various conditions like Adjacent to positive off-site feature

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?

Answer:

- A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data.
- The model should also be generalisable so that the test accuracy is not lesser than the training score.
- The model should be accurate for datasets other than the ones which were used during training.
- Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset.
- This would help increase the accuracy of the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.