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

The optimal value of alpha for ridge and lasso regression

Ridge Alpha 1

Lasso Alpha 10

Ridge(alpha=2)

r2_score_train - 0.882087717315285

r2_score_test - 0.8710808825348301

rss_train - 596084124320.2523

rss_test - 320797350989.88525

mse_train - 667507418.0517943

mse_test - 729084888.6133755

R2score on training data has decreased but it has increased on testing data

Lasso(alpha=20)

r2_score_train - 0.8854019697956436

r2_score_test - 0.8670105921065013

rss_train - 579329522996.7144

rss_test - 330925704432.2682

mse_train - 648745266.5136778

mse_test - 752103873.7097005

R2score of training data has decrease and it has increase on testing data

	Ridge2	Ridge	Lasso	Lasso20
LotArea	55922.640992	59778.431939	63955.064210	63617.887669
OverallQual	110944.014490	115599.252408	119957.483345	121719.072148
OverallCond	33226.593469	35638.745398	37354.981812	36948.765235
YearBuilt	54344.573607	54545.692314	53864.332906	53764.548095
BsmtFinSF1	52663.731203	51586.657410	50216.539701	50458.153814
TotalBsmtSF	74096.707724	76674.754264	78348.099735	78209.333502
1stFlrSF	71476.123090	73061.086063	8832.898863	8244.958141
2ndFlrSF	35224.759353	37149.879346	0.000000	0.000000
GrLivArea	85326.415089	87839.676484	163982.920640	162804.680303
BedroomAbvGr	-44604.715801	-52962.603870	-62831.358381	-61134.170375
TotRmsAbvGrd	53633.210113	52937.952456	51280.023696	50757.774874
Street_Pave	40419.432038	49959.412426	63045.460825	59515.001052
LandSlope_Sev	-21531.677392	-27846.862924	-37188.510825	-29661.614776
Condition2_PosN	-5843.960364	-11908.785655	-21920.323877	-11645.855795
RoofStyle_Shed	7274.217976	11641.731102	17801.452620	1966.058339
RoofMatl_Metal	11164.959608	18201.049929	32845.684073	16580.031007
Exterior1st_Stone	-23655.805061	-37132.047065	-69633.615929	-59674.587283
Exterior2nd_CBlock	-21223.133721	-32941.699298	-60463.906721	-49678.514531
ExterQual_Gd	-51867.902074	-54900.543840	-58459.152105	-57016.336034
ExterQual_TA	-60497.044122	-62317.508218	-64902.622534	-63508.829030
BsmtCond_Po	-4021.786999	-2488.039788	0.000000	-0.000000
KitchenQual_TA	-6282.925595	-5437.664855	-4495.491440	-4450.468043
Functional_Maj2	-15094.639225	-23574.925049	-40743.007254	-31654.783158
SaleType_CWD	-20812.381122	-27224.575631	-35460.118834	-30830.830798
SaleType_Con	16458.793758	21036.193759	25659.755739	21222.403113

LotArea----------Lot size in square feet

OverallQual---------Rates the overall material and finish of the house

OverallCond--------Rates the overall condition of the house

YearBuilt----------Original construction date

BsmtFinSF1--------Type 1 finished square feet

TotalBsmtSF------- Total square feet of basement area

GrLivArea----------Above grade (ground) living area square feet

TotRmsAbvGrd----Total rooms above grade (does not include bathrooms)

Street_Pave--------Pave Road access to property

RoofMatl_Metal----Roof material_Metal

Predictors are same but the coefficient of these predictors has changed

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

The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem

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

```
Index(['LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'BsmtFinSF1', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea',
'BedroomAbvGr', 'TotRmsAbvGrd', 'Street_Pave', 'LandSlope_Sev', 'Condition2_PosN', 'RoofStyle_Shed', 'RoofMatl_Metal', 'Exterio
r1st_Stone', 'Exterior2nd_CBlock', 'ExterQual_Gd', 'ExterQual_TA', 'BsmtCond_Po', 'KitchenQual_TA', 'Functional_Maj2', 'SaleTyp
e_CWD', 'SaleType_Con'], dtype='object')
```

LotArea,OverallQual,YearBuilt,BsmtFinSF1,TotalBsmtSF are the top 5 important predictors.

Let's drop these columns

```
X_train2 = X_train1.drop(['LotArea','OverallQual','YearBuilt','BsmtFinSF1','TotalBsmtSF'],axis=1)
X_test2 = X_test1.drop(['LotArea','OverallQual','YearBuilt','BsmtFinSF1','TotalBsmtSF'],axis=1)
```

```
# alpha 10
alpha =10
lasso21 = Lasso(alpha=alpha)
lasso21.fit(X_train2, y_train)
```

```
Lasso(alpha=10)
```

r2_score_train - 0.7988346707068132

r2_score_test - 0.7588103209258127

rss_train - 1016954777102.8658

rss_test - 600167078819.8167

mse_train - 1138807141.2126157

mse_test - 1364016088.226856

R2score of training and testing data has decreased

	Lasso21
OverallCond	7403.774043
1stFlrSF	163379.262938
2ndFlrSF	12227.759048
GrLivArea	186638.919740
BedroomAbvGr	-71218.036474
TotRmsAbvGrd	41610.305613
Street_Pave	101376.262107
LandSlope_Sev	-40205.679947
Condition2_PosN	0.000000
RoofStyle_Shed	53262.728685
RoofMatl_Metal	84219.173436
Exterior1st_Stone	-124162.644239
Exterior2nd_CBlock	-139534.253019
ExterQual_Gd	-77170.982079
ExterQual_TA	-108569.936019
BsmtCond_Po	-122646.594039
KitchenQual_TA	-11135.858324
Functional_Maj2	-48462.215856
SaleType_CWD	-64725.438438
SaleType_Con	52937.625483

Five most important predictor variables now are:

11stFlrSF----------First Floor square feet

GrLivArea----------Above grade (ground) living area square feet

Street_Pave---------Pave Road access to property

RoofMatl_Metal------Roof material_Metal

RoofStyle_Shed------Type of roof(Shed)

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

The model should be generalized 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 importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers 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. If the model is not robust, It cannot be trusted for predictive analysis.