Problem Statement -Part 2

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?

Ans:

Ans: The optimal value of alpha for ridge is

```
[368] print(ridge_model.best_params_)
```

```
{'alpha': 1.0}
```

The optimal value of alpha for lasso is

```
[369] print(lasso_model_cv.best_params_)
```

```
{'alpha': 10.0}
```

If we double the values of alpha then the Predictors are same but the coefficent of these predictor has changed

1 to 25 of 25 entries Filter

index	Variable	lasso_Coeff	ridge_Coeff	lasso_20_coff	ridge_2_coff
1	MSSubClass	63955.064	59778.432	63617.887738457015	55922.64099152115
2	LotFrontage	119957.483	115599.252	121719.07164636468	110944.0144903363
3	LotArea	37354.982	35638.745	36948.76515241326	33226.59346912354
4	OverallQual	53864.333	54545.692	53764.54797420014	54344.573607418046
5	OverallCond	50216.54	51586.657	50458.15381693855	52663.73120259231
6	YearBuilt	78348.1	76674.754	78209.33325956631	74096.70772440065
7	YearRemodAdd	8832.899	73061.086	8244.957998442693	71476.12308962205
8	MasVnrArea	0.0	37149.879	0.0	35224.759352940826
9	BsmtFinSF1	163982.921	87839.676	162804.68060189608	85326.41508859844
10	BsmtFinSF2	-62831.358	-52962.604	-61134.17022207664	-44604.71580077987
11	BsmtUnfSF	51280.024	52937.952	50757.77448575831	53633.21011302908
12	TotalBsmtSF	63045.461	49959.412	59515.00198127446	40419.43203756509
13	1stFlrSF	-37188.511	-27846.863	-29661.615180888202	-21531.67739208054
14	2ndFlrSF	-21920.324	-11908.786	-11645.85653126831	-5843.9603644890885
15	LowQualFinSF	17801.453	11641.731	1966.0587188136733	7274.217976074484
16	GrLivArea	32845.684	18201.05	16580.031429830768	11164.959608469464
17	BsmtFullBath	-60463.912	-32941.699	-49678.519872974386	-21223.133721125887
18	BsmtHalfBath	-69633.617	-37132.047	-59674.58801289473	-23655.80506080676

The most important predictor variables after the change is implemented are
LotArea,
OverallQual,
OverallCond,
YearBuilt,
BsmtFinSF1,
TotalBsmtSF,
GrLivArea,
TotRmsAbvGrd,
Street_Pave,
RoofMatl_Metal

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

Quetion 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: five most important predictor variables

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)

```
betas.head(68)
```

	Lasso_drop
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_CBlock	-139534.253019
Exterior1st_Stone	-124162.644239
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

Question 4

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: To make a model robust and generalisable 3 features are required

1.Model accuracy should >70-75%

2.P value of the features <0.05

3.IFA of all features <5

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.