Problem Statement - Part II

Q1. 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:

Optiomal Value of alpha for out housing model came out to be:

Ridge Regression - 10

Lasso Regression-100

On doubling the value of alpha

Ridge Regression

```
alpha = 10
ridge = Ridge(alpha=alpha)
# Fit the model on Training data
ridge.fit(X_train, y_train)
print(ridge.coef_)
```

```
[ 0.00000000e+00 -5.15774523e+03  2.34590552e+02
4.14924544e+03
  1.22453662e+04  6.25874347e+03  6.69113719e+03
3.09749123e+03
  1.30233872e+03 -5.52609886e+02 -2.06262756e+03
-5.45933663e+03
  1.36178744e+04 -2.04180141e+03  5.25059701e+03
0.00000000e+00
  1.91615364e+04  1.94971505e+03 -5.50090484e+02
```

-3.51445489e+02
  7.79151198e+02 -1.01848077e+03 0.00000000e+00
-1.57250234e+03
  1.27546034e+03 -2.06269301e+03 2.09080278e+03
5.42352546e+03
  1.80669058e+03 5.87395371e+02 1.47464768e+02
0.00000000e+00
  2.50849358e+02 0.00000000e+00 0.00000000e+00
5.08917370e+02
 -4.49242463e+02 6.01820658e+03 2.25742312e+03
3.46820623e+03
 -2.97650068e+03 -6.08528492e+02 -6.20169559e+03
1.44113219e+02
  5.79765044e+03 -3.93301192e+03 -1.59645707e+03
2.19449477e+03
  2.86270602e+02 2.38863337e+01 1.78149802e+03
4.96430549e+03
 -5.14521020e+03 1.56102198e+04 -1.34144517e+04
-3.22135574e+03
 -3.65250916e+02 -2.63560877e+03 -1.15211500e+04
-1.08290014e+04
  2.37354756e+03 -9.02170427e+03 1.44871162e+04
1.45935883e+04
 -7.27361118e+03 -1.77305939e+03 -9.17686935e+03
-3.79913744e+03
  1.13614918e+04 1.07938913e+04 2.61430716e+03
4.21431061e+03
  2.07171990e+03 3.43064686e+03 -9.61052771e+02
2.50030618e+02
 -2.59403669e+03 1.18650919e+03 3.43456703e+03
-1.82801420e+01
  4.05134468e+03 -1.26990144e+03 4.08860436e+03
-1.32635286e+03
 -8.96274735e+02 -2.34257993e+03 1.20852219e+04
-6.38301118e+02
  1.01014718e+02 -2.43599460e+03 -3.20949781e+02
-1.72439188e+03
 -3.85400059e+02 5.14206573e+02 6.87931696e+01
2.17225854e+02
 -6.66408798e+03 2.53828884e+03 -8.96274735e+02
-3.84715366e+02
  4.42071164e+03 -6.38301118e+02 5.08029073e+03
-2.39102715e+03
 -4.03819631e+03 1.28315632e+03 0.00000000e+00

```
-5.19031916e+03
  1.16637946e+03 -1.19815058e+03  2.18879366e+02
4.91348314e+03
 -5.33626363e+03  3.03048898e+03  2.16496303e+03
1.01846063e+03
 -1.75781029e+03  9.47513821e+02 -3.58274063e+03
2.69017825e+03
  2.90870420e+03 -5.91916394e+01  1.61443982e+03
-9.04733830e+02
 -4.55037597e+03 -1.15043360e+04 -4.30519075e+03
-8.87510558e+03
  6.50725147e+03 -1.42913764e+03 -4.75762804e+03
1.74286515e+03
  2.99777249e+03 -3.99967020e+03 -4.30519075e+03
-4.23508900e+03
 -1.61726941e+03  8.31851721e+02 -2.16749687e+03
-1.62858691e+03
 -1.80576455e+03 -7.58514504e+03 -7.48129587e+03
-8.32710438e+03
 -1.90225370e+03  2.33671199e+03 -2.77134058e+03
-1.18679183e+03
 -1.16859426e+02  3.90811629e+03  1.60104893e+03
4.49770068e+03
 -8.17320467e+03  4.00032276e+02  7.55457491e+02
7.55457491e+02
 -3.65280882e+03 -3.39438782e+03]
```

```
alpha = 20
ridge = Ridge(alpha=alpha)
# Fit the model on Training data
ridge.fit(X_train, y_train)
print(ridge.coef_)
```

```
[     0.          -4914.08467287    403.17836556
4009.98181375
  12894.2615192    6153.41445271   6791.91984316
3179.74639508
   1921.74427117   1588.73182569  -1337.91213154
-3378.59877441
  11643.27534811    245.60935152   5799.57899193
0.
  16594.57325059   1948.74583919   -621.67199658
-293.90757445
```

```
   1147.39394358  -1242.02937421      0.
-1122.07994488
   1894.40404915  -1948.20625782   2446.82856036
5231.41469271
   1856.5674806     749.32881123    184.22139294
0.
    304.21181269      0.              0.
548.89220111
   -472.28145049   4895.617778      907.27515848
2027.9676853
  -4122.42540462   -272.02365427  -3923.46807225
-254.05894491
   5060.57960197  -2876.86984255   -732.04821896
2076.24390514
    150.49210764    -68.28183354   1929.5924664
4001.26144917
  -4709.86862668  11901.07585467 -10434.32914069
-3130.81343133
    258.44180174   -721.73739301  -7972.07138823
-8314.17685276
   1526.07533089  -6608.4077176    9926.80413744
11282.7602589
  -5414.15901218   -644.11684809  -6354.14081521
-3223.59112749
   8273.52295142   6252.72338419   1764.53325677
2915.86423538
   1085.00475095   2021.35633947   -169.13170862
-292.81083308
  -1509.17011134   -181.96639237   1581.34972433
-335.87302036
   1955.51395471   -976.91835142   1689.71452228
-579.81182609
   -479.88813236  -1148.48643654   8865.99987668
-288.12794514
   1378.03753103  -2873.0262769    -131.52450844
-784.05478974
  -1422.80199209    468.3807758    -536.25244308
313.15389605
  -4597.81349083   1375.87360694   -479.88813236
-48.01073698
   3825.97691441   -288.12794514   3712.93440635
-1722.11654544
  -2145.60461647    756.94319435      0.
-3954.83814482
```

```
    891.66988831  -1072.50881763     63.87120095
2802.12029221
  -3700.95495147   1560.56547362   1975.16124905
332.1207147
  -1542.02835419   1096.61883011  -4432.10870901
1253.18122424
   2579.86401845   -206.36585356    661.25793438
-526.29427002
  -2430.50839095  -9073.56738963  -1898.22037322
-6538.88200461
   5854.79102608  -1386.69724119  -4806.78520424
1626.32382942
   3369.28076618  -2615.78054487  -1898.22037322
-3508.21133999
  -1176.77704195    534.88716373  -1889.57507252
-628.96908952
  -1850.77980821  -4511.51506022  -5680.51005079
-6387.05357526
  -1559.25142667   2528.51433014  -2449.48355598
-1441.96852752
   -596.87482258   2546.88807409    676.25084883
3527.38779046
  -5232.91118369   -693.46784002    573.57909684
x573.57909684
  -3450.97529523  -3189.38197681]
```

- **Coeff decreased for Ridge Regression -** In Ridge regression coeff are shrined towards zero. By shrinking the Coefficient towards zero, the model's variance will reduce leading to less complex model but could result in more biased data.

The most important variable after the changes has been implemented for ridge regression are as follows:-

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH

5. MSZoning_RM

6. SaleCondition_Partial

7. Neighborhood_StoneBr

8. GrLivArea

9. SaleCondition_Normal

10. Exterior1st_BrkFace

Lasso Regression

The value of alpha in Lasso regression is a hyperparameter that determines the strength of the regularization term in the model. A higher value of alpha means a stronger regularization term, which can lead to more parameters being set to zero and a simpler model. A lower value of alpha means a weaker regularization term, which can allow more parameters to be non-zero and result in a more complex model.

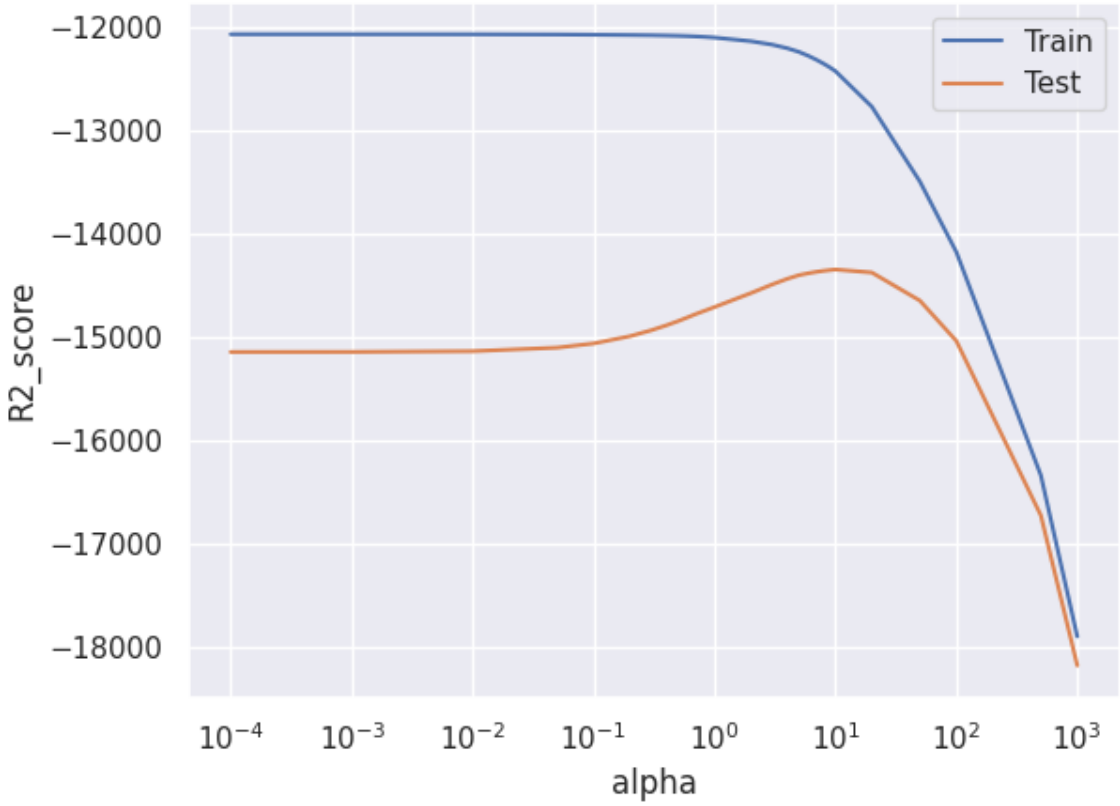
R2 score with alpha - 100

.9210212714443634

R2 score with alpha -200

.9141304509285384

-> R2 score value decreased - This means that model becomes more constrained and less able to explain the variance



Those 5 most important predictor variables that will be excluded are :-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

Q2. 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 model we will choose to apply will depend on the use case.
- If we have too many variables and one of our primary goal is feature selection, then we will use **Lasso**.
- If we don't want to get too large coefficients and reduction of

coefficient magnitude is one of our prime goals, then we will use **Ridge Regression**.

Q3. 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?

Ans:

Predictors with high coefficients or trait importance values are considered to be with most important predictors.

After dropping the important predictors cols from the data set and reach

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?

Ans

- Use large dataset so that we can have large train data and test data to train and model our data set to ensure its not overfitting the data.
- Handle Missing values effectively
- Validate the R2, VIF of the regression model.
- Regularize the model so that coefficients could be shrined towards zero and prevent he model from becoming too complex.
- If we look at it from the prespective of **Accuracy**, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.