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:

Ridged and lasso regression are two popular techniques used in linear regression to prevent from overfitting. It also addresses the problem of multicollinearity in the data. multicollinearity occurs when two or more independent variables in a regression model are highly correlated. The optimal value of alpha for ridged and lasso regression depends on the dataset and problem what you have doing.

Ridged Regression : If The Value Of Alpha Is High Then It Leads To Underfitting And If The Value Of Alpha Is Very Low Then It Leads To Overfitting. Ridge Regression Also Reduce The Magnitude Of Coefficient

Lasso Regression : For Lasso Regression The Alpha Value Is 1 The Output Is The Best Cross Validated Lambda Which Comes Out To Be 0.001 Once We Have The Optimal Value Of Lambda We Train The Lasso Model

To Find The Best Optimal Value Of Alpha For Lasso And Ridged Regression We Can Use Hyper Parameter Tuning And Doing Some Technique Like Cross Validation Techniques They Are Gridsearchcv And Randomsearch Cv

Changes In The Model If we Choose Double The Value Of Alpha For Both Ridge And Lasso:

By Changing The Value Of Alpha We Are Basically Controlling The Penalty Term. Higher The Value Of Alpha Higher The Penalty Term And Therefore The Magnitude Of Coefficient Are Reduced . If You Double The Value Of Alpha For Both Ridged And Lasso Regression The Model Will Become More Constraint And The Coefficient Will Be Smaller A Higher Value Of Alpha Means A Stronger Regularization Term, Which Can Lead To More Parameters Being Set To Zero And A Simpler Model

The Most Important Predictor Variables After The Change Is Implemented in ridged and Lasso regression:

Ridge regression

	variable	coeff
0	constant	10.991
3	OverallQual	0.463
10	GrLivArea	0.368
8	1stFlrSF	0.341
18	GarageArea	0.207
28	MSZoning_RL	0.160
15	BedroomAbvGr	0.149
13	FullBath	0.146
9	2ndFlrSF	0.142
26	MSZoning_FV	0.139
29	MSZoning_RM	0.139

Lasso Regression

	variable	coef
0	constant	10.816
10	GrLivArea	0.972
3	OverallQual	0.520
18	GarageArea	0.211
28	MSZoning_RL	0.202
26	MSZoning_FV	0.185
27	MSZoning_RH	0.179
29	MSZoning_RM	0.174
57	Neighborhood_NridgHt	0.142
11	BsmtFullBath	0.138
22	ScreenPorch	0.135

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 Lasso Regression Details :
-----------------------------------------------
the R2 score for train_lasso is: 0.8971683001148794
the R2 score for test_lasso is: 0.8008931994297012
the rss score for train_lasso is: 12.67921624384164
the Rss score for test_lasso is: 10.45541117957244
the mse score for train_lasso is: 0.012951191260308111
the mse score for test_lasso is: 0.024893836141839144
the rmse score for train_lasso is: 0.1138033007443462
the rmse score for t_lasso is: 0.15777780623978502
```

```
Ridge Regression metrics Details :
-----------------------------------------------
the R2 score for train_ridge is: 0.8968074572860429
the R2 score for test_ridge is: 0.8379249657753156
the rss score for train_ridge is: 12.723708402018229
the Rss score for test_ridge is: 8.51081489888162
the mse score for train_ridge is: 0.012996637795728528
the mse score for test_ridge is: 0.020263844997337188
the rmse score for train_ridge is: 0.11400279731536647
the rmse score for t_ridge is: 0.1423511327574782
```

Optimal Value Alpha :

- The optimal value of lambda (alpha) for **lasso regression** Is 0.0001.
- The Optimal Value Of Alpha For **Ridge Regression** is 2.0.

We Have Performed A Lasso And Ridge Regression. In Lasso Regression The R2 Score For Train Model Is 0.897 Whereas R2 Score For Ridge Regression Is 0.8968 Almost Both Of The Regression Are Performing Well Whereas In Ridge Regression The Train Score Is Slightly Low.

Similarly When We Compare The Both Regression Models The R2 Score For Test Is High In Ridge I.E 0.837 Whereas In Lasso Regression The R2 Score For Test Set Is 0.8000 Which Is Slightly Low When Compared To Ridge Regression. Hence Making The Model An Optimal Choice As It Seems To Perform In Unseen Data Well.

The MSE For Ridge And Lasso Model Perform Same On The Both The Training Data Set I.E 0.0129. And The MSE For Test Set (Ridge Regression) Lower Than That Of The Lasso Regression Model; Ridge Regression Performs Better On The Unseen Test Data Also Since Ridge Helps In Feature Selection

In Lasso Regression We Decrease The Alpha From 2,3,4 To 0.01 Then He R2 Square Value For Train Will Increase To 71% And For Test It Will Be The 64 %

Moreover While Choosing A Type Of Regression In The Real World An Analyst Has To Deal With The Lot Of Outliers ,Non-Normality Of Errors And Overfitting. When We Compare The Both The Models Both Are Performing Well In My Choice Better To Choose Lasso Regression Model

This Are The Some Metrics Which I Have Performed On The Model

	Metrics	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.015161e-01	0.896807	0.897168
1	R2 Score (Test)	-2.163877e+21	0.837925	0.800893
2	RSS(Train)	1.214313e+01	12.723708	12.679216
3	RSS(Test)	1.136286e+23	8.510815	10.455411
4	MSE(Train)	1.240360e-02	0.012997	0.012951
5	MSE(Test)	2.705443e+20	0.020264	0.024894
6	RMSE(Train)	1.113715e-01	0.114003	0.113803
7	RMSE(Test)	1.644823e+10	0.142351	0.157778

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:

Top 5 most predictor that will be excluded are :	***Top 5 predictor variables***:
1.GrLivArea	***1.Neighborhood_IDOTRR***
2.TotalBsmtSF	***2. BsmtCond_Gd***
3.OversllQual	***3. Neighborhood_NoRidge***
4.GarageArea	***4. LotArea***
5.OverallCond	***5. BedroomAbvGr***

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:

Robust & Generalizable :

A Model Is Said To Be Robust , When The Ability Of Model To Maintain Its Performance When Faced With Uncertainities This Includes Handling With Noisy Data .A Robust Model Should Be Able To Generalize Well And Provide Reliable Prediction Even When We Are Dealing With Unseen Data.

Generalization Is A Term Used To Describe A Models Ability To React On New Data That Is After Being Trained On A Training Set And Also Generalization Represents Its Ability To Make Accurate Predictions On New Data After Being Trained. Overfitting Occurs When A Model Is Well Trained And Find Exact Pattern On The Training Set And It Does Not Perform Well On Unseen Data I.E Test Data Or Making Inaccurate Prediction On New Data Similarly Underfitting Happens When A Model Is Insufficiently Trained And Does Not Have A Ability To Make Correct Predictions Even With The Training Data

There Is A Technique To Find Whether Our Model Is Robust Or Model Is Good Is Not That Is Bias-Variance Trade Off

If The Model Is Simple Then The Bias Is More And Variance Is Less Interms Of Accuracy Is That A Robust And Generalizable Model Will Perform Equally Well On Both Training And Test Data

The Bias Variance Trade-Off That Involves Finding The Right Balance Between Model Complexity And Its Ability To Generalize To New Unseen Data

Bias Varience Trade Off:

- ✓ **Bias** Means the Difference Between the Actual And Predicted
- ✓ Where As **Variance** Means How Predicted Values Are Scattered

The Trade-Off Occurs Because Increasing Model Complexity Then Bias Is Reduced Variance Is Increased, Vice-Versa. The Goal Is To Find The Optimal Level Of Complexity, To Minimize The Bias And Variance So That Model Can Perform Better On Unseen Data.

High Bias(Underfitting) : The Model Is To Simple And Does Not Learn The Pattern In The Data So It Can't Perform Well On Both Training And Test Data

High Variance (Overfitting) : The Model Is To Complex It Learn The Data Patterns Very Well Of Training Data And Create Noisy In The Data And Fails To Identify The Patterns On Unseen Data

Optimal Trade-Off: The Ideal Model Complexity Is Where The Total Error (Combination Of Bias And Variance) Is Minimized.