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 :

Ridge :

For the normal scanario optimum value of Alpha is 0.5 for Ridge. This is after doing the feature selection using RFE.

Without doing RFE feature selection , Alpha is 5.

Lasso :

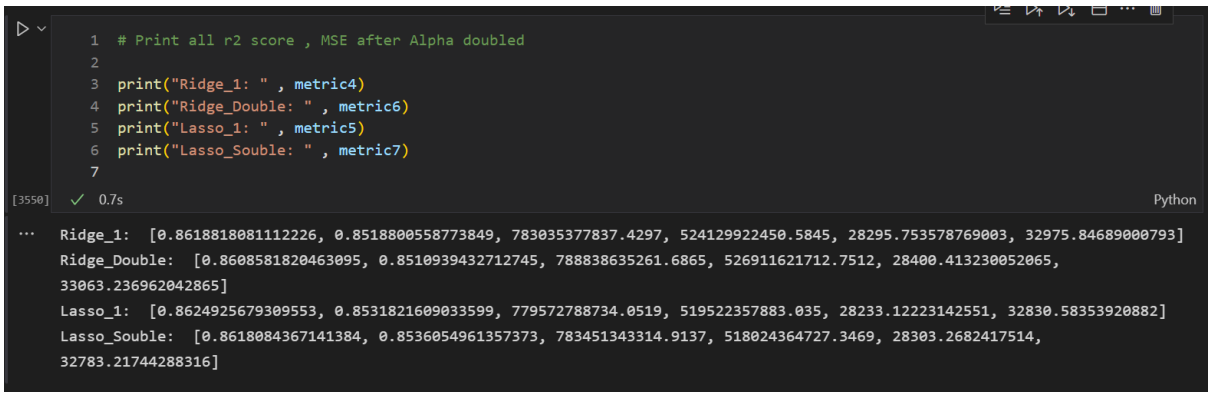
For the normal scanario optimum value of Alpha is 10 for Ridge. This is after doing the feature selection using RFE.

Without doing RFE feature selection , Alpha is 100.

After Doubling Alpha :

Ridge becomes .5 to 1 And Lasso Becomes 10 to 20

R2_Score , RSS and MSE after changing Alpha to double



```
# Print all r2 score , MSE after Alpha doubled

print("Ridge_1: " , metric4)
print("Ridge_Double: " , metric6)
print("Lasso_1: " , metric5)
print("Lasso_Souble: " , metric7)

[3550]  ✓ 0.7s                                                    Python

Ridge_1:  [0.8618818081112226, 0.8518800558773849, 783035377837.4297, 524129922450.5845, 28295.753578769003, 32975.84689000793]
Ridge_Double:  [0.8608581820463095, 0.8510939432712745, 788838635261.6865, 526911621712.7512, 28400.413230052065,
33063.236962042865]
Lasso_1:  [0.8624925679309553, 0.8531821609033599, 779572788734.0519, 519522357883.035, 28233.12223142551, 32830.58353920882]
Lasso_Souble:  [0.8618084367141384, 0.8536054961357373, 783451343314.9137, 518024364727.3469, 28303.2682417514,
32783.21744288316]
```

Changes Noticed :

1. r2_Score has slightly increased for test set after we changed Alpha to double
2. RSS and MSE value also slightly changed.But no drastic changes are there.
3. Feature contributing remains mostly same. But the Coefficient values has changed.

Most important features after Doubling Alpha :

GrLivArea	- Above grade (ground) living area square feet
OverallQual	- Rates the overall material and finish of the house
LotArea	- Lot size in square feet
GarageCars	- Size of garage in car capacity
OverallCond	- Rates the overall condition of the house
Old_Yr	- derived variable from when the house is build
MSSubClass	- Identifies the type of dwelling involved in the sale
GarageCExterior1stars	- Exterior covering on house
Neighborhood	- Physical locations within Ames city limits
TotRmsAbvGrd	- Total rooms above grade
BsmtFullBath	- Basement full bathrooms
FullBath	- Full bathrooms above grade
2ndFlrSF	- Second floor square feet
LotFrontage	- Linear feet of street connected to property

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:

R2_Score for Ridge : train - 0.861 Test - 0.851
R2_Score for Lasso : train - 0.862 Test - 0.853

R2 score is slightly higher for Lasso . Also the difference between Test and Train is also Slightly less in Lasso.
Also Lasso gives a slimpler model since it makes the less important feature value as 0.
Hence I will be choosing Lasso.

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 :

After removing the main 5 features , I could notice the below observations:

Alpha values has changed

R2_Score has decreased and RSS and MSE has increased.

Next Five Important features

- 1stFlrSF	- First Floor square feet
- 2ndFlrSF	- Second floor square feet
- TotRmsAbvGrd	- Total rooms above grade
- GarageCars	- Size of garage in car capacity/Type
- BsmtQual_Fa	- Basement Quality

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 :

Normally the model will be accurate when test accuracy and train accuracy are almost same value.

Most common issues faced by models are Multicollinearity, Overfitting , Non-constant variance, extrapolation and Autocorrelation .

To avoid these we need to take below mentioned steps :

1. To avoid Overfitting : Make sure not to train data completely on training set . So the accuracy of training set will be more. But test set will have less accuracy. To avoid this.

Make sure to eliminate the unnecessary features. Use sufficient and diversified training data .

2. To avoid Multi collinearty : Male sure to check the relation between variables and combine or remove related variables.
3. To avoid Non- Constant Variance : we need to transform response variables using Log or square roots.
4. To avoid Auto Correlation : Due to time dependent data. We need to try Auto Regression Model to avoid this.
5. Extrapolation : This happens when we try to predict values for which the model is not trained or outside training data. We can use diversified data to train model and also don’t predict values for which the model is not trained.
6. Outliers analysis should be done and Outliers should be removed Model shouldn’t be trained to accommodate all the outliers.which will lead to overfitting.