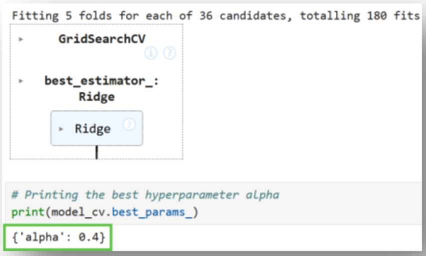
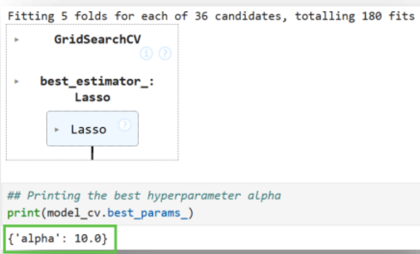
Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ridge – 0.4



Lasso – 10



With λ value given by model

With λ value as double

	Metric	Linear Regression	Ridge Regression (λ = 0.4)	Lasso Regression (λ = 10)	Ridge Regression (λ = 0.8)	Lasso Regression (λ = 20)
0	R2 Score (Train)	9.338634e-01	9.317449e-01	9.336823e-01	9.288546e-01	9.331390e-01
1	R2 Score (Test)	7.874006e-01	8.368718e-01	7.978182e-01	8.563885e-01	8.075843e-01
2	RSS (Train)	3.064113e+11	3.162264e+11	3.072501e+11	3.296169e+11	3.097673e+11
3	RSS (Test)	4.141704e+11	3.177941e+11	3.938755e+11	2.797732e+11	3.748499e+11
4	MSE (Train)	1.732365e+04	1.759893e+04	1.734735e+04	1.796767e+04	1.741826e+04
5	MSE (Test)	3.075052e+04	2.693617e+04	2.998765e+04	2.527353e+04	2.925443e+04

Top 5 Features - Ridge Regression (λ = 0.8)
GrLivArea
PoolQC_Gd
Condition2_PosN
OverallQual
YearBuilt

Top 5 Features - Lasso Regression (λ= 20)
PoolQC_Gd
Condition2_PosN
GrLivArea
OverallQual
YearBuilt
When lambda is doubled, top 5 features remain same but their coefficient values are different.

When lambda is doubled for Ridge and Lasso, R2 score (Train) remains approximately same. But there is slight increase in R2 Score (Test) for both Ridge and Lasso.

A slight increase in MSE(Train) but for MSE(Test) we see a slight decrease for Ridge model but not for Lasso.

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?

	Metric	Linear Regression	Ridge Regression (λ = 0.4)	Lasso Regression (λ = 10)
0	R2 Score (Train)	9.338634e-01	9.317449e-01	9.336823e-01
1	R2 Score (Test)	7.874006e-01	8.368718e-01	7.978182e-01
2	RSS (Train)	3.064113e+11	3.162264e+11	3.072501e+11
3	RSS (Test)	4.141704e+11	3.177941e+11	3.938755e+11
4	MSE (Train)	1.732365e+04	1.759893e+04	1.734735e+04
5	MSE (Test)	3.075052e+04	2.693617e+04	2.998765e+04

I choose Ridge Regression Model.

Although both models have similar R2 score on train data, R2 score is high on test data for Ridge Regression Model.

Also, MSE is higher for Lasso model on test data which shows model is deviating more for prediction of actual values.

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?

Top 5 features for Lasso

PoolQC_Gd
Condition2_PosN
GrLivArea
OverallQual
YearBuilt

```
LassoWTF.sort_values(by='Lasso Regression ($\lambda$ = 10) | without top features', ascending=False).head(5)
```

	Lasso Regression (λ = 10) | without top features
Condition2_PosA	102791.787246
GarageArea	69104.136342
Neighborhood_StoneBr	63706.620809
BsmtFinSF1	59202.652555
YearRemodAdd	54421.723265

```
LassoWTF.sort_values(by='Lasso Regression ($\lambda$ = 10) | without top features', ascending=True).head(5)
```

	Lasso Regression (λ = 10) | without top features
Condition1_RRAe	-28192.173972
Functional_Sev	-26217.338088
BldgType_Duplex	-25024.240990
BldgType_Twnhs	-18420.344195
Neighborhood_Mitchel	-14492.559787

Top 5 features after

Condition2_PosA
GarageArea
Neighborhood_StoneBr
BsmtFinSF1
YearRemodAdd

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?

A robust, generalizable model is required for

- performs consistently across unseen data,
- avoid overfitting models which works on train data very well but not on test data.

We need to make sure that R2 score of the model with test data data shold be reasonably good.

For a obverfiiting model, R2 score train data is ussually very high with very less R2 score for test data.

It can be achieved by

- Core Validation Strategy – K fold validation allows to use same train data to be used in different arrangements(folds) to be used to tarin the data and use rest of the data as test data.
- Regularization Techniques – To add penalties on features by shrinking coefficient values. This reduces overfitting of models.
  - Ridge – Add penalties by shrinking coefficient values close to zero but not zero
  - Lasso – Add penalties by shrinking coefficient values to zero
- Data cleanup
  - Handling outliers
  - Scaling of data
  - Removing corelated features using VIF