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?

Answer1:

Optimal value of alpha for lasso is 0.0001.

Optimal value for alpha for ridge is 0.8.

Model Metrics: Before Doubling the alpha

Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.835007	0.834605	0.834743
1	R2 Score (Test)	0.839343	0.841668	0.839739
2	RSS (Train)	4.784551	4.796222	4.792205
3	RSS (Test)	2.738889	2.699249	2.732143
4	MSE (Train)	0.068455	0.068539	0.068510
5	MSE (Test)	0.078987	0.078413	0.078890

Model Metrics After doubling the alpha

Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.835007	0.833710	0.834212
1	R2 Score (Test)	0.839343	0.842447	0.839562
2	RSS (Train)	4.784551	4.822167	4.807623

3	RSS (Test)	2.738889	2.685977	2.735156
4	MSE (Train)	0.068455	0.068724	0.068620
5	MSE (Test)	0.078987	0.078220	0.078933

```
Beta coefficients/slope value before doubling the alpha value
```

	Ridge	Lasso	lm
GrLivArea	0.269478	0.253386	0.255565
TotalSF	0.394402	0.433276	0.434690
house_age	-0.170639	-0.171868	-0.174527
MSZoning_RM	-0.048268	-0.046464	-0.047470
BldgType_Duplex	-0.065057	-0.064511	-0.067630
OverallCond_exce_veryexce	0.106830	0.105612	0.113590
BsmtQual_Ex	0.028714	0.026649	0.027153
BsmtExposure_Gd	0.027174	0.026088	0.027308
BsmtFinType1_ALQ	0.021241	0.019560	0.020308

	Ridge	Lasso	lm
BsmtFinType1_GLQ	0.031932	0.030171	0.029865
BsmtFinType1_Unf	-0.018300	-0.018791	-0.019310
Fireplaces_2	0.032169	0.030041	0.030777
FireplaceQu_Gd	0.029309	0.027961	0.027625
FireplaceQu_TA	0.022807	0.021386	0.021241
GarageType_Detchd	0.023294	0.022652	0.024529
GarageFinish_Fin	0.029739	0.028234	0.028113
GarageFinish_RFn	0.026393	0.024207	0.024502
GarageCars_3	0.029866	0.026978	0.026711
GarageCars_4	0.061545	0.031344	0.083249

Beta coefficients/slope value After doubling the alpha value

	Ridge	Lasso	lm
GrLivArea	0.274302	0.251142	0.255565
TotalSF	0.366617	0.431335	0.434690
house_age	-0.166799	-0.169498	-0.174527

	Ridge	Lasso	lm
MSZoning_RM	-0.048839	-0.045571	-0.047470
BldgType_Duplex	-0.062363	-0.061373	-0.067630
OverallCond_exce_veryexce	0.100990	0.097742	0.113590
BsmtQual_Ex	0.030095	0.026165	0.027153
BsmtExposure_Gd	0.026864	0.024885	0.027308
BsmtFinType1_ALQ	0.021813	0.018884	0.020308
BsmtFinType1_GLQ	0.033707	0.030519	0.029865
BsmtFinType1_Unf	-0.017296	-0.018256	-0.019310
Fireplaces_2	0.033331	0.029452	0.030777
FireplaceQu_Gd	0.030854	0.028226	0.027625
FireplaceQu_TA	0.024410	0.021452	0.021241
GarageType_Detchd	0.022191	0.020744	0.024529
GarageFinish_Fin	0.031258	0.028236	0.028113
GarageFinish_RFn	0.028062	0.023889	0.024502
GarageCars_3	0.032616	0.027332	0.026711
GarageCars_4	0.049218	0.000000	0.083249

Metrics such as R2 score for train and test, RSS value and MSE value has changed after doubling the alpha value. Value of Beta Coefficient tend to move towards zero after doubling.

The Top 5 most important predictors with their importance score a after the change has been implemented.

Attribute	Importance
TotalSF	**0.431335**
GrLivArea	**0.251142**
OverallCond_exce_veryexce	**0.097742**
BsmtFinType1_GLQ	**0.030519**
Fireplaces_2	**0.029452**

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?

Answer2.

I will choose to apply Lasso model because

1. **I have selected 171 features. There are too many variables. Hence lasso will be a better choice.**
2. **Lasso tends to move beta coefficient towards and eliminate the insignificant variables by making their beta value zero.**

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 3.

Five most important variables in lasso model is as under.

	Attribute	Importance
1	TotalSF	0.431335
0	GrLivArea	0.251142
5	OverallCond_exce_veryexce	0.097742
9	BsmtFinType1_GLQ	0.030519
11	Fireplaces_2	0.029452

If the above mention variables are not available then the top five important variable will be as follows in table below.

	Attribute	Importance
13	**GarageCars_4**	**0.114582**
12	**GarageCars_3**	**0.106327**
8	**FireplaceQu_TA**	**0.102080**

	Attribute	Importance
7	FireplaceQu_Gd	0.088725
3	BsmtQual_Ex	0.062500

Question 4

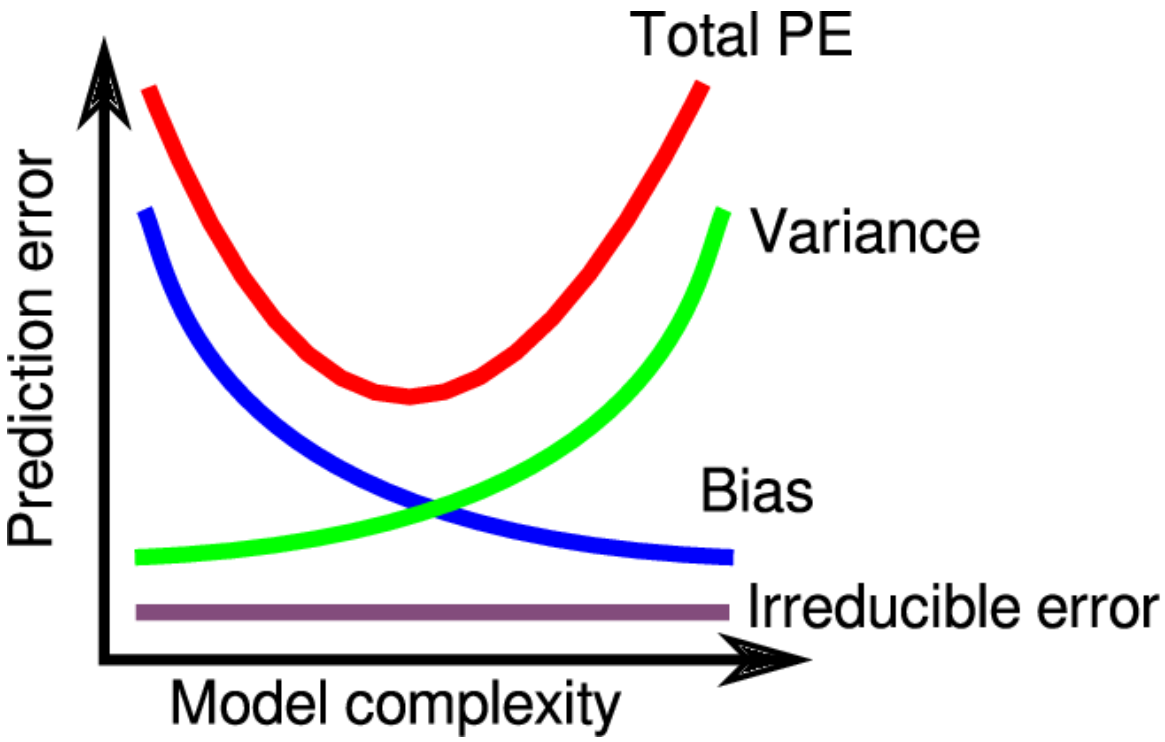
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Asnwer 4:

We can have more robust and generalizable model by choosing models that are simpler as they are more generic and tend to work better on unseen data. It would be advisable to avoid complex models as they lead to overfitting. Complex models will fit with high accuracy on training datasets but might fail miserably when applied on test data sets.

Regularization is an effective tool to improve the robustness and generalizability of the model.

If the model is robust and generalizable the accuracy score of test and and train will be very close. They might vary around +-5%.accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error. We need to find a optimal meeting point between bias and variance as shown in plot below to minimise the prediction error and keep the accuracy of the model high. Robust and generalizable models tend to have optimal bias-variance structure.


Total PE
Variance
Bias
Irreducible error
Prediction error
Model complexity