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 1:

The optimal value for alpha for Ridge regression is 0.5 and for Lasso regression is 0.0001.

The important predictor variables are listed in the table below

Feature	**Coefficient**
GrLivArea	0.320460
OverallQual	0.191716
OverallCond	0.108357
TotalBsmtSF	0.102850
LotArea	0.073387
HouseAgeInYears	-0.071191

Following is the table which shows the Ridge and Lasso metrics before and after it is doubled.

Metrics Name	**Linear**	**Ridge (alpha = 0.5)**	**Lasso (alpha = 0.0001)**	**Ridge (alpha = 1.0)**	**Lasso (alpha = 0.0002)**
r2ScoreTrain	0.905567	0.905288	0.904685	0.904643	0.902602
r2ScoreTest	0.887648	0.888057	0.888776	0.887611	0.887204
RSSTrain	1.706371	1.71142	1.722319	1.723075	1.759954
RSSTest	0.734098	0.731424	0.726729	0.734338	0.737002
MSETrain	0.040901	0.040962	0.041092	0.041101	0.041538
MSETest	0.040939	0.040865	0.040733	0.040946	0.04102

There is a reduction in r2score after the Alpha has been doubled for Lasso, but very minimal change in Ridge.

This is because the number of eliminated variables have increased when alpha has increased in Lasso, which has an minor effect on the model.

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 2

Ridge and Lambda imposes a penalty using a hyperparameter ‘lambda’. Ridge imposes penalty on sum of squared coefficients, whereas Lasso imposes penalty on sum of absolute values of the coefficients.

Since Ridge and Lasso has similar R2score Train and Test scores approx. 0.905 and 0.888 respectively, I would prefer to use Lasso, since it eliminates features and thus makes a simple model compared to Ridge.

Ridge doesn’t eliminate any features, but only shrinks the coefficients to nearly 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

After dropping the following significant predictor variables –

'GrLivArea', 'OverallQual', 'OverallCond', 'TotalBsmtSF', 'LotArea'

Following are the next five important predictor variables from Lasso regression with alpha 0.0001

Feature	**Coefficient**
TotalLivingAreaSF	0.486918
GarageCars	0.073396
FullBath	0.053718
CentralAir	0.045668
KitchenAbvGr	-0.042188

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 4:

A model can be considered robust when it performs well on unseen data.

Following steps can be followed to make the model robust.

- Ensure the quality and diversity in the data. The data collected needs to be cleaned, and validated to remove errors, outliers, missing values, and biases.
- Use various techniques to extract and create features, such as data transformation, aggregation, encoding, scaling, normalization, dimensionality reduction, feature selection.
- Use cross-validation, train-test split, or hold-out methods to validate models to avoid overfitting or underfitting.
- Compare and contrast different models based on their accuracy, precision, recall, f1-score depending on the usecase.
- Use regularization techniques, such as lasso, ridge, to reduce the complexity and variance of your models, and prevent overfitting.
- Test and deploy your models to see how they perform on new and unseen data, and in real-world situations.

Bias-Variance trade-off - The simpler the model, the more the bias but less variance and more generalizable.

If the model is robust and generic, then it can perform well on unseen data that differs from training data, i.e. the accuracy of the model on test data/real world data is almost same as training data.