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 regression:

Housing price prediction model built on ridge regression has following:

- Optimal value (alpha) = 3
- R^2 value of training dataset is 0.93
- R^2 value of testing dataset is 0.90

Most important predictor variables

Feature	Coefficient
Total SF	0.39
TotalBsmtSF	0.23
OverallQual_9	0.18
GarageArea	0.16
LotArea	0.16

After doubling the alpha for ridge regression has following

- Optimal value (alpha) = 6
- R^2 value of training dataset is 0.92
- R^2 value of testing dataset is 0.89

Most important predictor variables

Feature	Coefficient
Total SF	0.30
TotalBsmtSF	0.19
OverallQual_9	0.16
BsmtFinSF1	0.15
GarageArea	0.14

Housing price prediction model built on lasso regression has following

- Optimal value (alpha) = 0.001
- R^2 value for training dataset is 0.92
- R^2 value for testing dataset is 0.91

Most important predictor variables,

Feature	Coefficient
Total SF	0.75
TotalBsmtSF	0.28
OverallQual_9	0.22
GarageArea	0.18
OverallQual_10	0.17

After doubling the value of alpha for lasso regression has following

- Optimal value (alpha) = 0.002
- R^2 value for training dataset is 0.90
- R^2 value for testing dataset is 0.89

Most important predictor variables,

Feature	Coefficient
Total SF	0.74
TotalBsmtSF	0.27
GarageArea	0.19
OverallQual_9	0.19
BsmtFinSF1	0.12

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:

Ridge regression: Ridge regression penalises the loss function with the "squared magnitude" of the coefficient. For overfitting difficulties, this method is quite effective.

Lasso regression: The "absolute value of magnitude" of the coefficient is added to the loss function as a penalty term in Lasso Regression.

Although the overall accuracy of Ridge and Lasso is nearly identical, we might favour Lasso because the model we developed includes more features, including dummies.

In Lasso, the difference between train and test accuracy is minimal. As a result, Lasso is preferred over Ridge.

	Ridge regression	Lasso regression
Alpha value	3	0.001
R^2 value on training dataset	0.93	0.92
R^2 value on testing dataset	0.90	0.91

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:

Housing price prediction model built on lasso regression has following

- Optimal value (alpha) = 0.001
- R^2 value for training dataset is 0.92
- R^2 value for testing dataset is 0.91

Most important predictor variables,

Feature	Coefficient
Total SF	0.75
TotalBsmtSF	0.28
OverallQual_9	0.22
GarageArea	0.18
OverallQual_10	0.17

After removing top 5 features,

Housing price prediction model built on lasso regression has following

- Optimal value (alpha) = 0.0001
- R^2 value for training dataset is 0.92
- R^2 value for testing dataset is 0.88

After removing top 5 features the next top features with coefficients:

Feature	Coefficient
BsmtFinSF1	0.39
RoofStyle_Gable	0.36
TotRmsAbvGrd_9	0.34
RoofStyle_Hip	0.31
RoofStyle_Mansard	0.31

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:

Model is robust and generalisable when,

- If test data accuracy is not much lesser than the training data accuracy.
- The model should not be impacted by outliers.
- Predictor variables should be significant.

Implications for Accuracy of a model:

1. Gain more data as much as you can:

   It is beneficial to have more data since it allows the data to train itself rather than relying on weak correlations and assumptions.

2. Fix missing values and outliers:

   Missing values and outliers in the data can lead to an inaccurate model. Outliers can have an impact on the mean and median that we use to impute continuous variables.

3. Feature selection:

   It is solely based on domain expertise, allowing us to identify key features that have a significant impact on the target variable. The selection of features is also aided by data

visualisation. Significant variables can be found using statistical metrics such as p-Values and VIF.

4. Applying the right algorithm:

It is critical to select the appropriate machine learning algorithm in order to obtain an accurate model. This will come with experience.

5. Cross validation:

When more accuracy leads to overfitting, we can utilise the cross validation strategy, which involves leaving a sample on which the model was not trained and testing the model on this sample before moving on to the final model.