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

Ridge Model when using optimal alpha
RSS : 16.61
RMSE: 0.195
Most Important Predictor Variable : MSZoning_FV
MSZoning_FV denotes General zoning classification of the sale is Floating Village Residential

Ridge Model when we double the optimal alpha
RSS : 16.38
RMSE: 0.194
Most Important Predictor Variable : MSZoning_FV
MSZoning_FV denotes General zoning classification of the sale is Floating Village Residential

Lasso Model when using optimal alpha
RSS : 6.21
RMSE: 0.119
Most Important predictor Variable : GrLivArea
GrLivArea denotes above grade (ground) living area square feet

Lasso Model when we double the optimal alpha
RSS : 6.92
RMSE: 0.126
Most Important predictor Variable : GrLivArea
GrLivArea denotes above grade (ground) living area square feet

---

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

optimal value of lambda for ridge regression = *0.01*
optimal value of lambda for lasso regression = *0.001*

We obtained these optimal values by using

- GridSearchCV from sklearn.model_selection
- Kfolds used for cross validation = 5

- Scoring used to decide optimal values = neg_mean_squared_error
- List of alphas that we iterate over to get optimal value
    - lambdas = {'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1,
      0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
      4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000]}

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 predictor variables	**Coefficients (Betas)**
MSZoning_FV	*0.444308*
MSZoning_RL	*0.423775*
MSZoning_RH	*0.412456*
MSZoning_RM	*0.377497*
1stFlrSF	*0.297269*

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

To make a model robust and generalisable means to reduce complexity of model such that it gives good accuracy even on unseen data.Some of the steps we can follow are

1. Creating new features or updating existing features
    a. This can help in reducing multicollinearity or non linearity in data
    b. We can combine 3-4 features into a single feature which reduces the total number of features we have to work on
2. Using Log Transformation
    a. It can help in reducing the skewness or non-linearity in data which would increase accuracy of model

3. Using Recursive Feature Elimination(RFE)
    a. Before building Multiple Linear Regression model or Ridge Regression model we can use Feature Selection to reduce the number of features in our model which in turn reduces complexity of model
4. Regularization : Using Lasso Regression
    a. To reduce the complexity of model we want model coefficients to be as close to 0 as possible.
    b. This can be achieved with help of Regularization which adds a penalty term to Cost function
    c. We compromise a bit on the bias to get a significant reduction in the variance.
    d. Model complexity can be reduced by dropping unnecessary features
    e. Since Lasso Regression performs feature selection as well regularization it is the Linear Regression model when we have huge number of features