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:

- Optimal value of lambda for Ridge Regression is - 10.0
- Optimal value of lambda for Ridge Regression is - 0.001

If we double the value of alpha, following is the metrics for model:

Metrics	Ridge Regression	Lasso Regression
R2 Score (Train)	0.909757	0.916145
R2 Score (Test)	0.895220	0.900509
RSS (Train)	87.716390	81.507391
RSS (Test)	46.807666	44.445122
MSE (Train)	0.090243	0.083855
MSE (Test)	0.112249	0.106583
RMSE (Train)	0.300405	0.289578
RMSE (Test)	0.335035	0.326471

Changes in Ridge Regression Metrics:

- R2 Score for train dataset is decreased from 0.917 to 0.909
- R2 Score for test dataset is decreased from 0.898 to 0.895

Changes in Lasso Regression Metrics:

- R2 Score for train dataset is decreased from 0.924 to 0.916
- R2 Score for test dataset is decreased from 0.904 to 0.900

Top predictor variable after doubling the alpha value:

Ridge Regression	Lasso Regression
Neighborhood_StoneBr	Neighborhood_NridgHt
GrLivArea	OverallQual
Neighborhood_NridgHt	Neighborhood_NoRidge
Neighborhood_NoRidge	Neighborhood_StoneBr
SaleType_New	BsmtExposure_Gd

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:

Looking at metrics and coefficient **Lasso** is performing better in this particular case.

Selecting the model is depend on the use case:

- Lasso regression will perform better when the number of predictor variable are less as it is able to shrink variable completely 0 and remove them from model.
- Ridge regressions performs well when the number of predictor variables are more and their coefficients are likely equal. It keeps all the variable to the model.

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:

After dropping the top 5 predictor, Following are the new top 5 predictors -

- RoofMatl_WdShngl
- RoofStyle_Shed
- BsmtCond_Po
- RoofMatl_Membran
- SaleType_CWD

There is nothing much changed into the metrics after removing the top 5 variable.

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:

- A model is robust, if any small change into the dataset does not impact its overall performance.
- A generic mode should adaptive for new and should perform well on the unseen data with good accuracy.
- According to **Occam's razor**, A generic model shouldn't be more complex and it should be interpretable.
- There is a trade-off between "Too Simple" and "Too Complex: Model:
    - If model is too simple, it leads to underfitting, it means it does not learn all the important patterns and leading to low accuracy (**Bias**).
    - If model id too complex, it leads to overfitting and it does not generalise well on the unseen data. This leads to performance of the model (**Variance**) on any change in the data.
    - This variation is called Bias, Variance trade off.
    - We should find the optimal balance in the model using Regularization techniques like Ridge and Lasso Regression.

See the below diagram:

Overfitting