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 for alpha are:

Lasso Regression model : 0.001
Ridge Regression model: 4.0

Their respective metrices :

Current model Metrics:	
For Ridge Regression Model (Original Model, alpha=4.0): For Train Set: R2 score: 0.9118105765808643 MSE score: 0.08818942341913573 MAE score: 0.21137950782035478 RMSE score: 0.2969670409643732 For Test Set: R2 score: 0.8910807896472409 MSE score: 0.13247384381707972 MAE score: 0.24817794428693607 RMSE score: 0.3639695644103772	**For Lasso Regression Model (Original Model: alpha=0.0001):** For Train Set: R2 score: 0.9101962206246228 MSE score: 0.08980377937537722 MAE score: 0.21311795404307757 RMSE score: 0.29967278717857787 For Test Set: R2 score: 0.892248540621506 MSE score: 0.13105355753625722 MAE score: 0.2482628336706969 RMSE score: 0.36201320077623855
After doubling the Alpha values :	
For Ridge Regression Model (Doubled alpha model, alpha=4*2=8): For Train Set: R2 score: 0.910032805950911 MSE score: 0.08996719404908904 MAE score: 0.21303867809885157 RMSE score: 0.29994531843169187 For Test Set: R2 score: 0.8915493583060288 MSE score: 0.1319039435109647 MAE score: 0.248351648445895 RMSE score: 0.3631858250413481	**For Lasso Regression Model: (Doubled alpha model: alpha:0.001*2 = 0.002)** For Train Set: R2 score: 0.9055879880365295 MSE score: 0.09441201196347054 MAE score: 0.21802473817508525 RMSE score: 0.30726537709847906 For Test Set: R2 score: 0.8929278124772253 MSE score: 0.13022738781438306 MAE score: 0.25031768533727317 RMSE score: 0.3608703199410878

Observation:

- A slight increase in test R2 score is observed when alpha values are doubled for both lasso and ridge models
- A slight increase in test RMSE score is observed when alpha values are doubled for both lasso and ridge models
- For train set the R2 and RMSE scores are slightly better than the newly build model .

Current model Metrics:	
For Ridge Regression (Doubled alpha model, alpha=4*2=8): The most important top10 predictor variables after the change is implemented are as follows: ['GrLivArea', 'MSZoning_FV', 'MSSubClass_90', 'AgeofProperty', 'Neighborhood_Crawfor', 'MSSubClass_160', 'Exterior1st_BrkComm', 'Neighborhood_StoneBr', 'Neighborhood_NridgHt', 'OverallQual']	**For Lasso Regression (Doubled alpha model: alpha:0.001*2 = 0.002**): The most important top10 predictor variables after the change is implemented are as follows: ['GrLivArea', 'MSZoning_FV', 'Neighborhood_Crawfor', 'Exterior1st_BrkComm', 'AgeofProperty', 'MSSubClass_90', 'MSSubClass_160', 'Neighborhood_StoneBr', 'Neighborhood_NridgHt', 'OverallQual']
After doubling the Alpha values :	
For Ridge Regression (Doubled alpha model, alpha=4*2=8): The most important top10 predictor variables after the change is implemented are as follows:	**For Lasso Regression (Doubled alpha model: alpha:0.001*2 = 0.002**): The most important top10 predictor variables after the change is implemented are as follows: ['GrLivArea', 'Neighborhood_Crawfor', 'AgeofProperty', 'MSZoning_FV',

['GrLivArea', 'MSZoning_FV', 'AgeofProperty', 'Neighborhood_Crawfor', 'MSSubClass_90', 'MSSubClass_160', 'OverallQual', 'Neighborhood_NridgHt', 'Neighborhood_StoneBr', 'Exterior1st_BrkFace']	'OverallQual', 'MSSubClass_160', 'Neighborhood_NridgHt', 'MSSubClass_90', 'Neighborhood_StoneBr', 'Exterior1st_BrkFace']

Observation:

- There is a minor change in top 10 pedictor of model and there alignment too.
- 'Exterior1st_BrkComm' predictor is missing in new model for both lasso and ridge(when alpha is doubled).
- Rest predictors are same.

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:

As we have seen that the metrics are almost similar for lasso and ridge in the above question,but lasso regression model gives slightly better R2 score and RMSE value for test set i.e unseen data. So we should use lasso regression model.

One more advantage we can have in lasso that we can do feature elimination too using lasso .

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:

As per the jupyter notebook,after removing first 5 most important predictor we will get below predictors as first 5:

['Foundation_Slab', 'Neighborhood_MeadowV', 'Neighborhood_IDOTRR', 'MSSubClass_30', 'Neighborhood_OldTown', 'Exterior1st_BrkFace']

Question 4

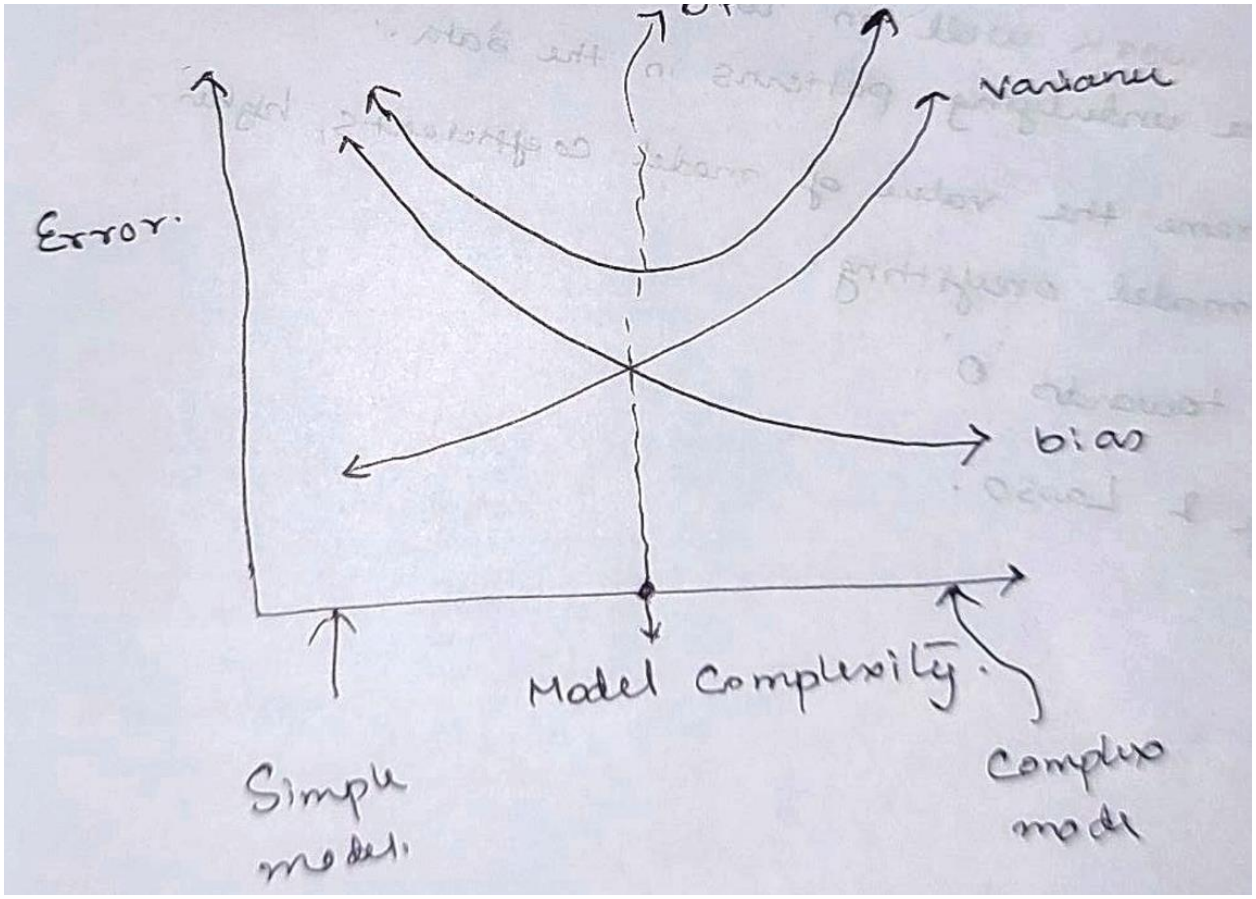
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.



So there should be a tradeoff between bias and variance to obtain a model with minimum error.The diagram shows optimal model complexity at the intersection of bias and model as well as minimum error.