PROBLEM STATEMENT - PART II

(Subjective Questions)

Q.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?

=>

a) The optimal value of alpha for ridge and lasso regression are -

1) Ridge Regression - 0.0001
2) Lasso Regression - 50

b) The changes in the model after doubling the value of alpha for both ridge and lasso are as follows -

Initially,

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.271299e-01	9.271299e-01	9.231552e-01
1	R2 Score (Test)	9.015324e-01	9.015342e-01	9.039920e-01
2	RSS (Train)	3.639745e+11	3.639745e+11	3.838272e+11
3	RSS (Test)	2.237036e+11	2.236996e+11	2.181158e+11
4	RMSE (Train)	2.071751e+04	2.071751e+04	2.127502e+04
5	RMSE (Test)	2.313387e+04	2.313366e+04	2.284312e+04

After doubling the alpha,

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.271299e-01	9.271299e-01	9.206644e-01
1	R2 Score (Test)	9.015324e-01	9.015359e-01	9.037927e-01
2	RSS (Train)	3.639745e+11	3.639745e+11	3.962687e+11
3	RSS (Test)	2.237036e+11	2.236955e+11	2.185684e+11
4	RMSE (Train)	2.071751e+04	2.071751e+04	2.161708e+04
5	RMSE (Test)	2.313387e+04	2.313346e+04	2.286681e+04

Inference - After doubling the values of alpha, Ridge Regression is not affected as such, but the Lasso Regression has been affected a bit. The R2_score has reduced and the error metrics have increased by a small amount for Lasso Regression.

c) The most important predictor variables after the change is implemented are as follows -

Ridge Regression

1) GrLivArea
2) BsmtFinSF1
3) BsmtUnfSF
4) Exterior1st_BrkComm
5) NoYearBuilt
6) RoofStyle_Shed
7) RoofStyle_Gambrel
8) OverallQual
9) RoofStyle_Hip
10) RoofStyle_Gable

Lasso Regression

1) GrLivArea
2) BsmtFinSF1
3) OverallQual
4) NoYearBuilt
5) BsmtUnfSF
6) OverallCond
7) GarageArea
8) Neighborhood_StoneBr
9) Neighborhood_NoRidge
10) KitchenQual_TA

Q.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?

=>

We will choose to apply Lasso Regression as compared to Ridge Regression, because

- The r2_score of lasso regression is slightly higher than ridge regression for the test dataset.
  Ridge (r2_score) - 0.9015341521693286
  Lasso (r2_score) - 0.9039919558214556

- Also the error values are slightly less.
  Ridge (rss) - 223699555421.2732
  Lasso (rss) - 218115796215.34952
  Ridge (rmse) - 23133.663793903503
  Lasso (rmse) - 22843.119859535476

- And most importantly, Lasso Regression helped in feature selection as well.

Q.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?

=>

The five most important predictor variables in the original lasso model are:-

1) GrLivArea
2) BsmtFinSF1
3) OverallQual
4) NoYearBuilt
5) BsmtUnfSF

After removing these attributes from the dataset, we again build the lasso model. The new model had the following metrics:-

r2_score (Train) - 0.8526319196155496
r2_score (Test) - 0.8045400336259317
rss (Train) - 736079757462.563
rss (Test) - 444055563871.5201
rmse (Train) - 29462.155079243603
rmse (Test) - 32593.46383648255

The five most important predictor variables now are -

1) 1stFlrSF
2) GarageArea
3) BedroomAbvGr
4) Neighborhood_StoneBr
5) Neighborhood_NoRidge

Q.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?

=>

A model is considered to be robust if its output and predictions are consistently accurate even if one or more of the input variables or assumptions are drastically changed due to unforeseen circumstances.

According to the principle of Occam's razor, a model should be as simple as possible as well as robust. A simpler model has the following advantages:-

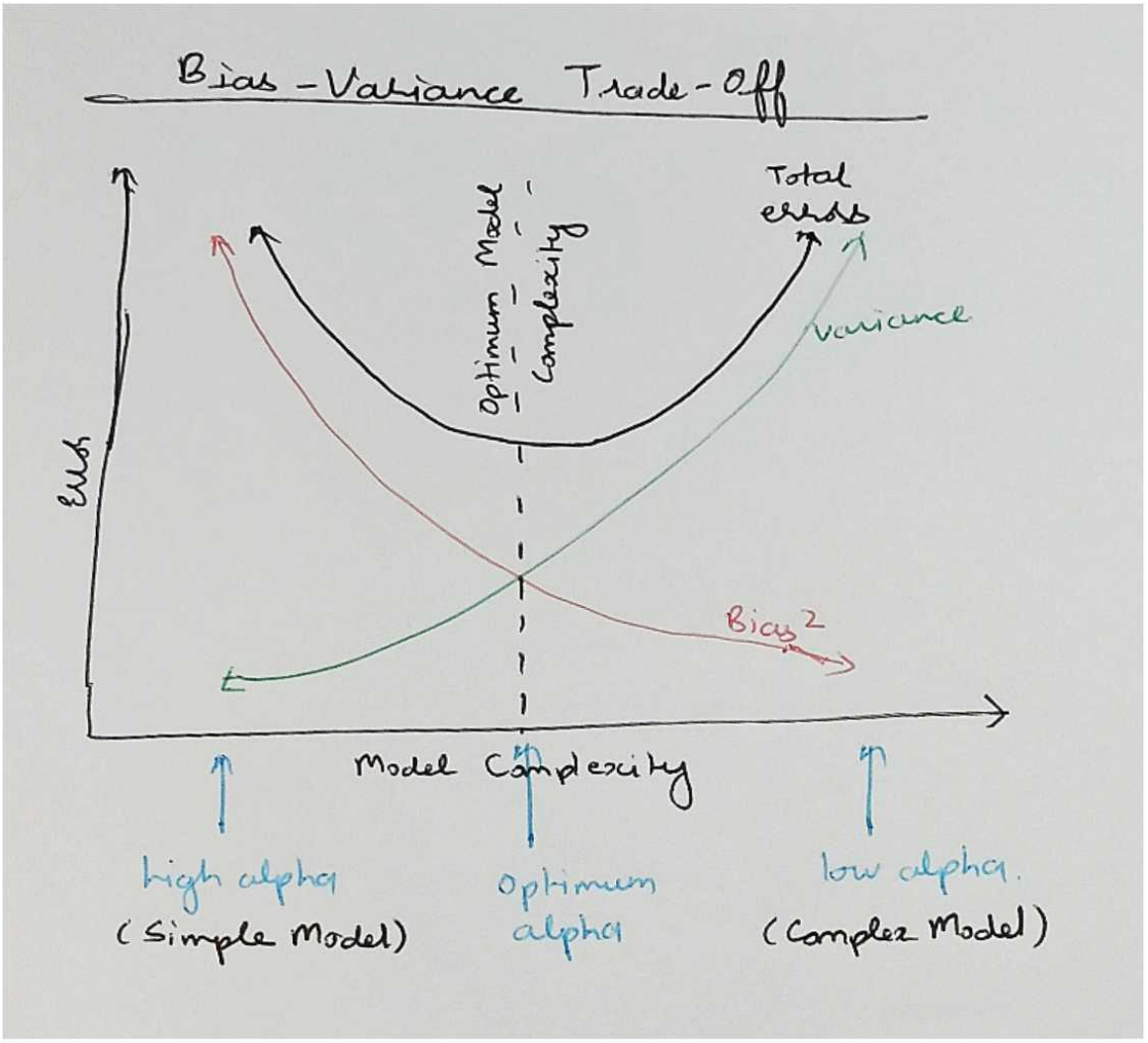
1) Simpler models are usually more generic
2) Simpler models require fewer training examples
3) Simpler models are more robust
    - Complex models tend to change wildly with changes in the training data
    - Simpler models do not change significantly if the training data points undergo small changes.

Regularization helps in making the model simpler by penalizing the coefficients. But simplicity has its own disadvantages. Extremely simple models are likely to fail in complex real world phenomena.

Thus, to make sure that a model is robust and generalizable, one needs to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use. This leads us to the bias variance trade-off.

1) Bias quantifies how accurate the model is likely to be on unseen data
2) Variance refers to the changes in the model as whole when trained on a different data set.

Thus, accuracy of a model can be maintained by keeping the appropriate balance between bias and variance as it minimizes the total error as shown in graph below:-



Therefore, accuracy and robustness may be at odds with each other as too much accurate model may fall prey to overfitting and thus may fail on unseen data.