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:

The optimal value for ridge is 100 and for lasso is 500 and double the optimal value for ridge is 200 and lasso is 1000.

Description	Metrics	Ridge	RidgeDouble	Lasso	Lasso Double
Train	r2	0.905	0.89	0.907	0.88
	mse	483628710.24	526413427.33	470713061.48	570139259.94
	rmse	21991.55	22943.70	21695.92	23877.58
Test	r2	0.839	0.83	0.834	0.82
	mse	1282706857.96	1311451811.92	1323193872.36	1366063271.68
	rmse	35814.89	36213.97	36375.73	36960.29

Based on the data, when we change the alpha value to double there is 1% of decrease to the r2 score and it doesn't affect the metrics

The top 5 most important features of rigid regression after the implementation are **OverallQual, GrLivArea, NridgHt, 1stFlrSF, GarageCars**

The top 5 most important features of lasso regression after the implementation are **GrLivArea , OverallQual, NridgHt, GarageCars, StoneBr**

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:

The Optimal value for Ridge is 100 and Optimal value for Lasso is 500.

Based on the metric table

Description	Metrics	Ridge	RidgeDouble	Lasso	Lasso Double
Train	r2	0.905	0.89	0.907	0.88
	mse	483628710.24	526413427.33	470713061.48	570139259.94
	rmse	21991.55	22943.70	21695.92	23877.58
Test	r2	0.839	0.83	0.834	0.82
	mse	1282706857.96	1311451811.92	1323193872.36	1366063271.68
	rmse	35814.89	36213.97	36375.73	36960.29

Due to larger number of feature, it better to use lasso so it will negligible the minimum coefficient with zero and metric wise lasso seems to quite equal or better than ridge model. So its better to choose lasso 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:

The top 5 most important features of lasso regression are **GrLivArea, CompShg, OverallQual, WdShngl, NridgHt** .

The removal of these features got drop in r2 score in lasso model from 0.907 to 0.879 and 3% reduce in r2 score.

The newly trained model 5 important features are **1stFlrSF, 2ndFlrSF, YearBuilt, TotRmsAbvGrd, MasVnrArea**

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?

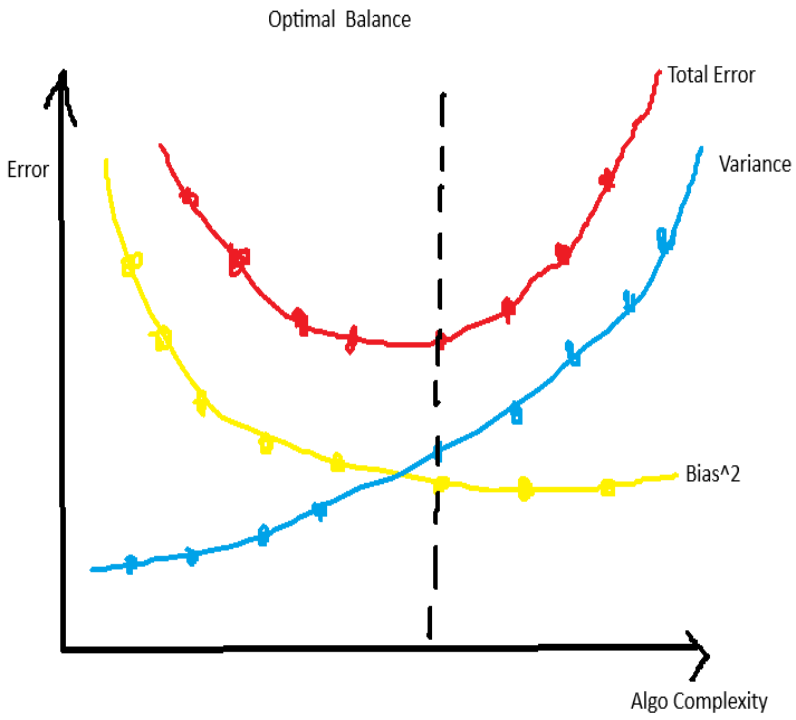
Answer:

Occam's Razor

The model should be simple as possible so that the model is generalized, robustness, required less data for learning and only few assumptions. Its better to choose simple model to complex one.

Bias-Variance Tradeoff

- Bias is the difference between our actual and predicted values. Bias is the simple assumptions that our model makes about our data to be able to predict new data.
- Variance as the model's sensitivity to fluctuations in the data.
- As complexity increases, the bias reduces and variance increases and vice versa, The main objective is to find the optimal value of total error. An optimal balance of bias and variance will never overfit or underfit the model
    - Total error = bias error + variance error



Overfitting

The model that memories the complete data on the training set and it does not preform well on the test set is called overfitting. This will affect the model to be generalized.

This can prevent by the following steps

1. Adding more training data
2. Regularization
3. Simplify the data
4. Cross Validation
5. Feature selection etc.