House Price Prediction Assignment subjective questions

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 is 5 and 100 respectively.

- The R2 Scores that we got are:
  - With Ridge of alpha 5:
    - → For Train Set it is: 0.9536964080483873
    - → For Test Set it is: 0.8835914704189932
  - **With Lasso for alpha 100:**
    - → For Train Set It is: 0.9438503004089455
    - → For Test Set it is: 0.8901859603368314

- Then if the Alpha value was doubled:
  For Ridge it is 10 and for Lasso it is 200
  - With Ridge of alpha 10:
    - → For Train Set it is: 0.9467159463880558
    - → For Test set it is: 0.8818476728969056
  - With Lasso for alpha 200:
    - → For Train set it is 0.9319133093254242
    - → For Test set it is: 0.886503891900433

So, by this we can conclude that I have got the Quite similar results in the R2 Score with slight change in the score by changing the alpha values to the double of previous one.

The important variables after the change implementation are:

KitchenAbvGr	LotArea
OverallCond_3	RoofMatl_WdShngl
KitchenQual_Fa	Neighbourhood_StoneBr
BsmtQual_TA	SaleType_New
KitchenQual_TA	GarageCars
KitchenQual_Gd	OverallQual_8
BsmtQual_Gd	OverallQual_5
OverallCond_4	BsmtQual_Fa
OverallQual_9	OverallQual_10
GrLivArea	Total SF

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?

   A. The values of lambda here are referring to the alpha values that we got previously they are 5, 100 and after doubling the alpha values it is 10, 200 for both lasso and ridge regressions. And based upon this we can say that Lasso is better option and it helps in some of the features for feature elimination.

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?

   A. After excluding the most important features we got to conclusion by seeing at the histogram we can say that there are some top variables that help in good predictors they are: Total SF, GrLivArea, LotArea, RoofMatl, SaleType, GarageCars, etc.

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. 1. The model will be robust and generalizable when:

   - The test accuracy is not much lesser than the training accuracy score.
   - The predicted variables should be significant.
   - Model Significance can be determined by p-value, R2 score and the adjusted R2.
   - Always a simple model can be robust in nature.
   - And last but not the least the model should not be impacted by outliers.

2. The implications of Accuracy of a model:

- We have to gain as much more data as we can. This helps in making the model learn the good data to perform well on the test data to predict better values on unseen data.
- Fix the missing values and also the outliers in the data, so that the data reads the perfect dataset to train and test the models in regression of code.
- Feature Engineering or newly derived columns we have to standardize the values using Scaler feature in the code
- Feature Selection play a major role in building a model this helps in acquiring the business problem and helps in solving the criteria further. After building a good model by good feature selection the perfect insights will be given by the model to identify the variables that affect the business domain the most.
- Applying the right algorithm also plays a crucial role in building the model. This helps the model to get the right accuracy.
- Cross validation also helps in getting a good fit model. Sometimes the data fit is good but the model may underfit so for that we use cross validation to resolve the problem of overfitting or underfitting o f a model.