Housing-Price-Model-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?

- The optimal value of alpha for ridge and lasso regression model is 20 and 0.0001 respectively.
- If the alpha value of both the model gets doubled there are change in the coefficients as shown below:

```
Ridge (alpha=20.0)  Lasso (alpha=0.0001)  Ridge (alpha = 40.0)  Lasso (alpha = 0.0002)
MSSubClass  -0.017490 -0.017679 -0.016908 -0.017207
LotArea 0.037675  0.037415  0.037881  0.037490
LandSlope 0.012426  0.012651  0.012176  0.012532
OverallQual 0.084335  0.086599  0.082375  0.086806
OverallCond 0.041810  0.042738  0.040908  0.042772
BsmtQual  0.022614  0.021076  0.023809  0.021138
BsmtExposure  0.015316  0.014792  0.015646  0.014685
BsmtUnfSF -0.025217 -0.027912 -0.022868 -0.027831
TotalBsmtSF 0.090564  0.095333  0.086619  0.095284
HeatingQC 0.009744  0.009215  0.010169  0.009235
CentralAir  0.015992  0.015601  0.016326  0.015577
2ndFlrSF  0.050335  0.053600  0.047573  0.053186
BsmtFullBath  0.016870  0.015385  0.017967  0.015318
FullBath  0.029693  0.028194  0.030658  0.028176
HalfBath  0.017816  0.016447  0.018753  0.016408
KitchenQual 0.015490  0.013993  0.016875  0.014106
TotRmsAbvGrd  0.044546  0.044455  0.044416  0.044311
Functional  -0.019632 -0.019873 -0.019279 -0.019721
Fireplaces  0.020847  0.020633  0.021111  0.020734
FireplaceQu 0.013318  0.012369  0.014194  0.012412
GarageArea  0.040321  0.040046  0.040601  0.040196
GarageQual  0.011152  0.011254  0.010905  0.011039
YearBuilt_Old -0.029797 -0.032492 -0.027619 -0.032074
YearRemodAdd_Old  -0.023037 -0.023430 -0.022640 -0.023200
GarageYrBlt_Old 0.016191  0.016774  0.015335  0.016285
MSZoning_RM -0.021440 -0.021652 -0.021221 -0.021579
LotConfig_CulDSac 0.012314  0.012196  0.012322  0.012087
Neighborhood_CollgCr  -0.011256 -0.011295 -0.011026 -0.011111
Neighborhood_Edwards  -0.016019 -0.015834 -0.016058 -0.015627
```

```
Neighborhood_Gilbert  -0.013209 -0.012942 -0.013221 -0.012733
Condition1_Norm 0.019256  0.019452  0.018945  0.019355
Condition2_Norm 0.009997  0.010192  0.009752  0.010117
BldgType_TwnhsE 0.013919  0.014470  0.013096  0.014106
Exterior1st_HdBoard -0.022315 -0.025993 -0.018704 -0.023831
Exterior1st_Plywood -0.015348 -0.017337 -0.013246 -0.016196
Exterior1st_VinylSd -0.008489 -0.008670 -0.008129 -0.008397
Exterior1st_Wd Sdng -0.024959 -0.027444 -0.022294 -0.026330
Exterior2nd_HdBoard 0.010529  0.013815  0.007349  0.011795
Exterior2nd_Plywood 0.008581  0.010720  0.006393  0.009483
Exterior2nd_Wd Sdng 0.018744  0.021203  0.016176  0.020079
MasVnrType_BrkFace  0.009413  0.012274  0.006442  0.006636
MasVnrType_None 0.008839  0.012334  0.005263  0.006438
MasVnrType_Stone  0.009559  0.011077  0.007935  0.007532
Foundation_CBlock -0.011658 -0.012321 -0.011124 -0.012193
Foundation_PConc  0.018364  0.017743  0.018613  0.017716
PavedDrive_Y  0.009157  0.008573  0.009600  0.008599
SaleCondition_Normal  0.014783  0.015191  0.014206  0.015071
SaleCondition_Partial 0.018410  0.018987  0.017686  0.018824
```

- After changes are implemented the most important variables remained the same as before, as shown below:

```
Ridge (alpha=20.0)  Lasso (alpha=0.0001)  Ridge (alpha = 40.0)  Lasso (alpha = 0.0002)
TotalBsmtSF 0.090564  0.095333  0.086619  0.095284
OverallQual 0.084335  0.086599  0.082375  0.086806
2ndFlrSF  0.050335  0.053600  0.047573  0.053186
TotRmsAbvGrd  0.044546  0.044455  0.044416  0.044311
OverallCond 0.041810  0.042738  0.040908  0.042772
```

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?

- After inspecting the model I decided to go with lasso regression as the R2 score and RMSE value for test was better compare to Ridge regression. And after observing the coefficients for both of the regressions model, Lasso model showed better values.

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 most important 5 features after the modifications are:

```
            Lasso coefficents
1. FullBath    0.088605
2. LotArea     0.062131
3. GarageArea  0.057726
4. HalfBath    0.057493
5. BsmtQual    0.057487
```

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?

- I have perfomed the model evaluation and I verified the assumptions required for the regression models which are

    1. Linearity: Relationship between the dependent variable and the independent variables were linear.
    2. Homoscedasticity: The variance of the residuals were seen to be constant across all the levels of the independent variables.
    3. Normality: The residuals were normally distributed.
    4. No Autocorrelation: There were no visible pattern seen for the residuals, which tells the independence of errors.