Australian house price prediction

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?

What is the optimal value of alpha for ridge and lasso regression?

- Optimal value of alpha for Ridge regression is 10.
- Optimal value of alpha for Lasso regression is 0.0001.

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso ?

➢ If we are changing the alpha value of Ridge regression to 20 then following changes are taking place-:

- R2 score for train data is decreasing by 0.0027
- R2 score for test data is increasing by 0.0004
- RSS for train data is increasing by 0.40
- RSS for test data is decreasing by 0.03
- MSE for train data is increasing by 0.001
- MSE for test data is staying equal up to 3rd precision
- RMSE for train data is increasing by 0.001
- RMSE for test data is is staying equal

➢ If we are changing the alpha value of Lasso regression to 0.0002 then following changes are taking place-:

- R2 score for train data is decreasing by 0.0003
- R2 score for test data is increasing by 0.001
- RSS for train data is increasing by 0.004
- RSS for test data is decreasing by 0.07
- MSE for train data is staying equal
- MSE for test data is staying equal
- RMSE for train data is staying equal
- RMSE for test data is staying equal

Observation for above analysis is the R2 score for train data is going down when we are increasing the value of regularization . And the value of RSS is going up for train data set as the bias is increasing. While for test data or unseen data the R2 score is increasing and the RSS is going down as the variance is decreasing.

What will be the most important predictor variables after the change is implemented?

- Top features after changes implemented to Ridge regression are shown in below image.

```
[('GrLivArea', 0.1288022004096922 9),
 ('OverallQual', 0.11009585379052444),
 ('TotalBsmtSF', 0.097006423127704 79),
 ('YearRemodAdd', 0.069730292892824 47),
 ('Exterior1st_VinylSd', 0.06367918745275183),
 ('Exterior1st_BrkFace', 0.06052593638621 1024),
 ('Neighborhood_Crawfor', 0.0585156984156856),
 ('Neighborhood_Somerst', 0.05357601485345618),
 ('FireplaceQu_Gd', 0.0514850458619061),
 ('SaleCondition_Normal', 0.04337973755598019),
 ('LandContour_Low', 0.041336136411322),
 ('Exterior1st_Plywood', 0.03906779439614654),
 ('Exterior2nd_CmentBd', 0.03082617601640 7806),
 ('Exterior2nd_MetalSd', 0.02675447800732282),
 ('Exterior1st_HdBoard', 0.02510088734227 2987),
 ('Exterior1st_WdShing', 0.02032644863137 2302)]
```

- Top features after changes implemented to Lasso regression are shown in below image.

```
[('GrLivArea', 0.1309157041597706),
 ('Exterior1st_BrkFace', 0.11112180018328224),
 ('OverallQual', 0.10420211270230903),
 ('TotalBsmtSF', 0.09506941970498337),
 ('Neighborhood_Crawfor', 0.08954209763389011),
 ('Exterior1st_VinylSd', 0.08856667878973365),
 ('Neighborhood_Somerst', 0.07158678409652011),
 ('Exterior2nd_CmentBd', 0.06845979624557258),
 ('YearRemodAdd', 0.06771950834988588),
 ('Exterior1st_Plywood', 0.06362403481262328),
 ('LandContour_Low', 0.062446772077881),
 ('FireplaceQu_Gd', 0.05184798669241658),
 ('SaleCondition_Normal', 0.047609422074981034),
 ('Exterior1st_WdShing', 0.04726388583574496),
 ('Exterior1st_HdBoard', 0.04703112036491797),
 ('Exterior2nd_MetalSd', 0.046223019949880 33)]
```

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?

	LinearRegression	Ridge	Lasso
R2Score(train)	0.885801	0.883443	0.885710
R2Score(test)	0.843058	0.846919	0.844218
RSS(train)	16.677649	17.022011	16.690969
RSS(test)	10.044615	9.797476	9.970386
MSE(train)	0.016319	0.016656	0.016332
MSE(test)	0.016319	0.022369	0.022763
RMSE(train)	0.127744	0.129057	0.127795
RMSE(test)	0.151436	0.149562	0.150876

According to the evaluation metrics found in our assignment above we can say

- R2 score of Ridge regression on test data set is slightly higher than the Lasso regression.
- RSS of test data set for Ridge regression is lower than the lasso regression.
- MSE of test data set for Ridge regression is lower than the lasso regression.
- RMSE of test data set for Ridge regression is lower than the lasso regression.

Thus we are choosing the Ridge regression for regularization .

Question 3

After building the model, you realized 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?

```
[('FireplaceQu_Gd', 0.1956083917747687),
 ('Exterior2nd_CmentBd', 0.1649834679147250),
 ('YearRemodAdd', 0.1408992843985397),
 ('LandContour_Low', 0.0965124152684236),
 ('Exterior1st_Plywood', 0.0927283017847213)]
```

Top 5 features after removal of top 5 features and rebuilding lasso regression model using optimal alpha .i.e. 0.0001 are shown in the above picture.

Question 4

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

- The model will be called a robust model when adding more input data will not change the variance so much. And there exists a balanced variance bias trade off.
- And it will be generalizable when the model performs better .i.e. the r2 score and adjusted r2 score of testing data is not so far off from the r2 score and adjusted r2 score of training data.
- The accuracy of the model can be explained as the ability to perform better on unseen data , and reducing the Residuals sum of squares , which means the model should explain most of the underlying pattern and also give similar best performance for unseen data.
- A robust model would not induce more error when added more input data , and the generalizable model performs equally better in terms of explainability in both seen and unseen data set .
- This implicates the fact that a robust and generalized model will be more accurate for any business problem .