Housing Assignment-Subjective-Questions

Q1. 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?

ANS: Following are the optimal values for both:

- Optimal value of lambda for Ridge Regression = 10
- Optimal value of lambda for Lasso = 0.001

Changes in Ridge Regression metrics:

R2 score of train set decreased from 0.8754074992627734 to 0.8624353055188951

R2 score of test set decreased from 0.8587679649562701 to 0.8503619039007676

Changes in Lasso metrics:

R2 score of train set decreased from 0.8432555208853181 to 0.7936328185290887

R2 score of test set decreased from 0.8337194426845449 to 0.7897105316042708

Note: In both cases there are no changes in predictor variables, please get the following screenshot of the predictor variable.

Ridge:

	Params	Coef
0	constant	0.160
11	GrLivArea	0.073
9	2ndFlrSF	0.063
8	1stFlrSF	0.058
20	GarageCars	0.058
18	TotRmsAbvGrd	0.057
82	BsmtQual_Ex	0.051
14	FullBath	0.051
3	MasVnrArea	0.040
38	Neighborhood_Crawfor	0.036
86	BsmtExposure_Gd	0.034

Lasso:

	Params	Coef
11	GrLivArea	0.333
0	constant	0.200
20	GarageCars	0.096
82	BsmtQual_Ex	0.065
86	BsmtExposure_Gd	0.029
57	OverallQual_8	0.029
43	Neighborhood_NridgHt	0.023
91	BsmtFinType1_GLQ	0.021
14	FullBath	0.020
3	MasVnrArea	0.018
115	MSSubClass_20	0.017

Q2. 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?

ANS: As per the final data model for housing pricing, ridge is better than Lasso due to the following reasons.

1. Ridge R2's score is better than Lasso R2's score.
2. Ridge alpha value is 10 and Lasso alpha value is 0.001 because Lasso value is too small.
3. Lasso tends to do well if there are a small number of significant parameters and the others are close to zero but, in our case, there are more than 80 significant parameters so ridge will perform better.

Q3. 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?

ANS: Here, we will drop the top 5 features in the Lasso model and build the model again. ***Top 5 Lasso predictors were:***

8	1stFlrSF	0.205
9	2ndFlrSF	0.118
19	GarageArea	0.089
13	FullBath	0.041
17	TotRmsAbvGrd	0.039

Top 5 Ridge predictors were:

11	FullBath	0.082
16	GarageArea	0.065
77	BsmtQual_Ex	0.057
13	BedroomAbvGr	0.053
3	MasVnrArea	0.049

Q4. 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?

ANS: A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.

Cross validation to check if the model is robust or not, we split the data into multiple subsets and apply the same model on all subsets then we take the average of all coefficients.

Hyperparameter tuning: To find the optimal value that balance the tradeoff between bias and variance