Submitted by:
Shubhanshu Kumar Singh
IIITB Roll No: EML21100082
Applicant ID: APFE21709564

Advanced Regression
Assignment Part 2- Subjective Questions

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:

- Optimal value of alpha for Ridge is 1.0
- Optimal value of alpha for Lasso is 0.0001

R2 Score **before** we **double** the values of alpha:

- R2 Score (Train):
  - Ridge Regression: **0.840298**
  - Lasso Regression: **0.840199**
- R2 Score (Test):
  - Ridge Regression: **0.839711**
  - Lasso Regression: **0.840889**

R2 Score **after** we **double** the values of alpha:

- R2 Score (Train):
  - Ridge Regression: **0.838767**
  - Lasso Regression: **0.839152**
- R2 Score (Test):
  - Ridge Regression: **0.838242**
  - Lasso Regression: **0.840816**

We can see that after we double the values of alpha, R2 score drops for both Ridge and Lasso regression on train and test set. Also Lasso penalizes and removes one of

the features known as "**Exterior1st_Stone**". According to data dictionary this depicts Exterior covering stone on the house.

After we double the values of alpha, we get the following top 5 predictors (Sorted Lasso column coefficients in descending order):

	Feature	Ridge	Lasso
0	OverallQual_Very Excellent	0.203227	0.221557
1	TotalBsmtSF	0.167703	0.184122
2	OverallQual_Excellent	0.164144	0.170478
3	2ndFlrSF	0.100434	0.104895
4	OverallQual_Very Good	0.096258	0.097541

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:
Based on the final result, we can see that Train R2 score for Linear, Ridge and Lasso Regression are almost same but Test R2 score is slightly higher for Lasso.
We'll consider Lasso as our final model because we know it helps in feature selection. It also reduces overfitting.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.840945	0.840298	0.840199
1	R2 Score (Test)	0.840300	0.839711	0.840889
2	RSS (Train)	1.915096	1.922877	1.924070
3	RSS (Test)	0.776483	0.779344	0.773618
4	MSE (Train)	0.043609	0.043698	0.043711
5	MSE (Test)	0.042396	0.042474	0.042318

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:
Final Metric after dropping 5 features and training the model on remaining 9 features:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.637938	0.629586	0.623372
1	R2 Score (Test)	0.627044	0.620981	0.625494
2	RSS (Train)	4.359388	4.459950	4.534767
3	RSS (Test)	1.813361	1.842841	1.820894
4	MSE (Train)	0.065796	0.066550	0.067106
5	MSE (Test)	0.064789	0.065313	0.064923

We can see a large drop in R2 scores.

After sorting the Lasso column coefficients in descending order, we get the following top 5 predictors:

	Feature	Linear	Ridge	Lasso
0	GarageCars	0.203447	0.181556	0.203911
1	FullBath	0.173786	0.149038	0.154405
2	BsmtFinSF1	0.136060	0.120391	0.116916
3	Fireplaces	0.103369	0.102878	0.101507
4	LotFrontage	0.109452	0.085686	0.054472

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:

A model is robust and generalisable when it can capture the variance in unseen/test datasets very well. A model with low bias and low variance is generally preferrable. Its accuracy on test set should be as good as seen on training set. While model building, we need to take care of EDA and data cleaning part. Like handling the outliers, imputation of missing values, removing the highly skewed features (excluding target feature), handling multicollinearity and make sure that the assumptions of a particular algorithm are being followed. Like in case of Linear Regression algorithm we need to check whether all the assumptions are correct for the dataset or not. Only after that we can apply that particular algorithm for our use case. In case of Regression problem, we can select relevant features using Lasso and RFE rather than picking up all independent features.