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?

Answer:

The optimal value for alpha is
1. Ridge = **0.7**
2. Lasso = **0.0001**

After we double the optimal value for alpha we can see slight changes in scores mentioned below. Once we double the value we see slight decrease in **R2_Score** and slight increase in **RSS** and **RMSE** score.

	Ridge		**Lasso**	
Alpha	**0.7**	**1.4**	**0.0001**	**0.0002**
R2 Score (Train)	0.900825	0.898964	0.897521	0.889077
R2 Score (Test)	0.876795	0.875718	0.879663	0.873312
RSS (Train)	1.684164	1.715757	1.740271	1.883662
RSS (Test)	0.939513	0.947720	0.917637	0.966068
RMSE (Train)	0.001650	0.001680	0.001704	0.001845
RMSE (Test)	0.002145	0.002164	0.002095	0.002206

Similarly we also note change in the predictors mentioned in below tables for Ridge and Lasso marked in yellow.

Ridge	
When Alpha = 0.7	**When Alpha = 1.4**
OverallQual	OverallQual
1stFlrSF	1stFlrSF
GrLivArea	GrLivArea
MSZoning_RL	MSZoning_RL
MSZoning_RH	MSZoning_RH
MSZoning_FV	GarageArea
LotArea	OverallCond
GarageArea	MSZoning_FV
OverallCond	LotArea
MSZoning_RM	BedroomAbvGr

Lasso	
When Alpha = 0.0001	**When Alpha = 0.0002**
GrLivArea	GrLivArea
OverallQual	OverallQual
MSZoning_RL	GarageArea
MSZoning_RH	OverallCond
GarageArea	FullBath
OverallCond	1stFlrSF
MSZoning_FV	BedroomAbvGr
MSZoning_RM	BsmtFullBath

1stFlrSF	Neighborhood_Crawfor
BedroomAbvGr	BsmtQual

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 :

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.902029	0.900825	0.897521
1	R2 Score (Test)	0.876515	0.876795	0.879663
2	RSS (Train)	1.663720	1.684164	1.740270
3	RSS (Test)	0.941648	0.939513	0.917636
4	RMSE (Train)	0.001630	0.001650	0.001704
5	RMSE (Test)	0.002150	0.002145	0.002095

The **R2-Score** of **Lasso** is slightly higher than **Ridge** for the test dataset so we will **choose Lasso Regression** to solve this problem.

Also Lasso helps in Feature elimination by bringing some of the coefficients to 0.

Q3. 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?

Ans:

The five most important predictor variables in the lasso model are:

1. **GrLivArea**
2. **OverallQual**
3. **MSZoning_RL**
4. **MSZoning_RH**
5. **GarageArea**

After dropping above mentioned variable we get other set of five most important variable:

1. **1stFlrSF**
2. **2ndFlrSF**
3. **OverallCond**
4. **BsmtQual**
5. **FullBath**

Q4. 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?

Ans:

1. Perform a grid search over a range of λ values to systematically explore different regularization strengths and their impact on the model.

2. Use cross-validation techniques, such as k-fold cross-validation, to find the optimal value for the regularization hyperparameter (λ).

3. Split the dataset into training and validation sets multiple times and train the model with different values of λ. Choose the value that provides the best performance on the validation set.

4. Understand that a small value of λ will result in a model similar to a standard linear regression, while a large value will lead to more regularization.

5. Balancing regularization strength is crucial.

6. Cross-validation, grid search, and evaluating on a test set is crucial for making the model robust and generalizable.

7. **Implications of the accuracy**: The accuracy of the model will go up if we try to over fit the model but that no longer makes it generalizable. When the model is generalized the accuracy should be pretty good on both the training and the testing dataset making the model robust.