Advanced Regression 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?

Solution:

The optimal values of alpha for lasso and ridge model are as below.

1. **Ridge:** Optimal value of alpha for ridge model is 0.5
2. **Lasso:** Optimal value of alpha for lasso model is 0.0001.

Considering the Optimal value for Ridge and Lasso below are the top 5 features

Ridge top 5 predictors with their coeffiecients are :

Features	Coefficients
GrLivArea	0.3222
OverallQual	0.2054
OverallCond	0.1056
TotalBsmtSF	0.1056
Functional	0.0672

Lasso top 5 predictors with their Coefficients are:

Features	Coefficients
GrLivArea	0.3308
OverallQual	0.2108
OverallCond	0.1057
TotalBsmtSF	0.1049
Functional	0.0637

After doubling the values for alpha, for ridge and lasso. i.e., Ridge = 1 and Lasso = 0.0002 below are the top 5 features for Ridge and Lasso models along with their coefficient values.

Ridge top 5 predictors with their coefficients are :

Features	Coefficients
GrLivArea	0.305171
OverallQual	0.204071
YearBuilt	0.137284
TotalBsmtSF	0.105974
OverallCond	0.100251

Lasso top 5 predictors with their Coefficients are:

Features	Coefficients
GrLivArea	0.3308
OverallQual	0.216155
YearBuilt	0.138691
TotalBsmtSF	0.104584
OverallCond	0.099586

Changes in the model if you choose double the value of alpha for both ridge and lasso:

- **Lasso regression** shrinks coefficients all the way to zero, thus removing them from the model.
- **Ridge regression** shrinks coefficients toward zero, but they rarely reach zero. Coefficients of predictors decrease, and then their value in the model decreases. That is, their effect decreases. And thus the flexibility of the model should decrease.

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:

We consider below points while considering the regularization models:

- If there are too many variables and if want to reduce the complexity then we can opt for Lasso model.
- If the variables are important in business point of view and you need it in your model then use Ridge model.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.919211	0.918959	0.918884
1	R2 Score (Test)	0.889109	0.889099	0.890041
2	RSS (Train)	1.316858	1.320959	1.322187
3	RSS (Test)	0.802669	0.802741	0.795921
4	MSE (Train)	0.038640	0.038700	0.038718
5	MSE (Test)	0.046081	0.046083	0.045887

From the above table we can see that R2_scores are almost same for both ridge and lasso but as lasso will penalize more on the dataset and can also help in feature elimination thus I will choose Lasso as final model. Also the given dataset is large, where we have around 120 features + (some of which are obsolete and needs to be removed)

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:

So after excluding these 5 features-

- GrLivArea
- OverallQual
- YearBuilt
- TotalBsmtSF
- OverallCond

\\

Five most important predictor variables now are :

	feature	importance
6	GarageCars	0.259207
8	OpenPorchSF	0.149553
2	BsmtFinSF1	0.144440
1	YearRemodAdd	0.137517
0	LotArea	0.132951

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:

- The model should 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.
- A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data.
- That outlier which 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.
- This would help standardize the predictions made by the mode. If the model is not robust, it cannot be trusted for predictive analysis.
- To make sure that a model is robust, model must be built using Median or Median Absolute Deviation which are unaffected by outliers in the data. And will not be affected while predicting on unseen or altering dataset.

Hence, we should always select a model which is just complex enough to understand the variance in the data without much inaccuracy at the same time not too complex to over fit.
This can be achieved using regularization. Regularization is the process of deliberately simplifying models to achieve the correct balance between keeping the model simple and yet not too naïve.