Assignment: Advanced Regression-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: The Optimal Value for Alpha for Ridge and Lasso came out to be 4 and 0.0005 respectively. This value was when the target variable was not scaled. On doubling the values of Alpha, results are as below:

Initial Values			Final Values (On Doubling Alpha)		
Ridge (Alpha =4):			**Ridge (Alpha= 8)**:		
R2 Score (Train)		0.91165	R2_train Score:		0.905606
R2 Score (Test)		0.89238	R2_test Score:		0.886116
Top 5 Features:			Top 5 Features:		
1	OverallQual	0.316395	1	OverallQual	0.272845
2	GrLivArea	0.257206	2	GrLivArea	0.203055
3	1stFlrSF	0.226597	3	OverallCond	0.18444
4	OverallCond	0.225263	4	1stFlrSF	0.175308
5	TotRmsAbvGrd	0.170152	5	TotRmsAbvGrd	0.169151
Lasso (Alpha = 0.0005):			**Lasso (Alpha = Alpha= 0.001):**		
R2 Score (Train)		0.906216	R2_train Score: 0.8961340171752619		0.896134
R2 Score (Test)		0.899161	R2_test Score: 0.8922483011061014		0.892248
Top 5 Features:			Top 5 Features:		
1	GrLivArea	0.885173	1	GrLivArea	0.785663
2	OverallQual	0.49991	2	OverallQual	0.573998
3	OverallCond	0.283481	3	OverallCond	0.233007
4	GarageCars	0.221291	4	GarageCars	0.222338
5	BsmtFullBath	0.143573	5	BsmtFullBath	0.132346

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: Lasso would be the preferred option over the Ridge Regression even though the Ridge has a higher R2 Score. This is because Lasso helps to perform feature selection. as for a few features the coefficient value is exactly zero hence making the model is less complex.

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: When we drop the top 5 features from Lasso Model, we again get the alpha value as 0.0002 on Lasso. The details are as below:

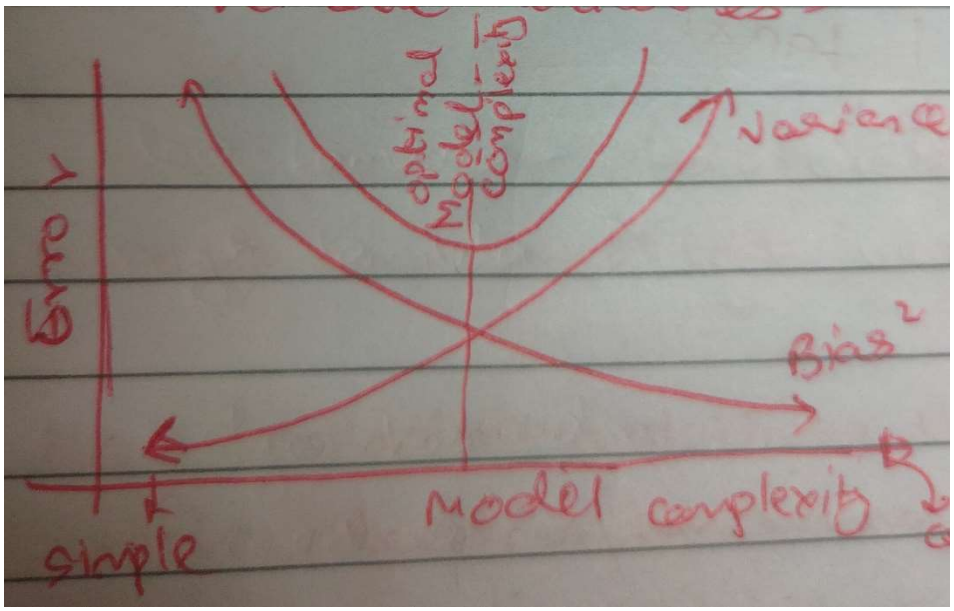
Lasso (Alpha = 0.0002):		
R2_train Score:		0.903351
R2_test Score:		0.871076
	Top 5 Features	
1	1stFlrSF	0.854207
2	2ndFlrSF	0.408671
3	MSZoning_FV	0.33994
4	MSZoning_RL	0.298723
5	MSZoning_RH	0.274798

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 robust model is one whose output dependent variable is accurate even when the Independent variables change drastically. However, an accurate model is not possible due to **bias and variance** that creeps into a model. To make a balance model we need to ensure we have a balance between bias and Variance.

As seen in image below, when we have high bias the variance is very low, the model is too simple and the model would be underfit. However, with low bias, we have high variance hence model turns to be overfit. We need to take case of this and try to find a balance between bias and Variance.



In Regression what **metric** to use is quite important, instead of r2_score, we could focus more on Adjusted r2_score, as it explains how well the selected independent variable explains the variance of dependent variable. The model should be generalized so that the test accuracy is not too low than the train score.

Another important aspect is to see how **stable** the model is. We would want every time we use the model, we get same performance. While building the model always go for cross validation as it helps in validating the model even during the training phase.

Model should not be too **sensitive** to outliers. Model should be stable enough to outliers or noise in the data. Model should perform well in such scenarios.

When new data comes in model should perform, in almost the similar manner as it did during the model development phase.