Advanced Regression Assignment (Problem Statement – Part II)

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

From the python models, the optimal alpha value for ridge is **6.59** and for lasso is **0.0003.**

When the alpha values are doubled, the model suffers from reduction in R-square for both training and test set.

	Ridge Regression			Lasso Regression		
	Optimal Alpha (α = 6.59)	Double Alpha (α = 13.18)	Delta	Optimal Alpha (α = 0.0003)	Double Alpha (α = 0.0006)	Delta
Train R2	0.8895	0.8499	~ -0.04	0.8776	0.8590	~ -0.018
Test R2	0.8668	0.8324	~ -0.033	0.8575	0.8436	~ -0.014

It's observed that for Ridge Regression, Training Set R2 decreased by 4% pts while Test Set R2 decreased by 3.3% pts.

In Lasso, the decrease was less – Training Set R2 suffered drop in R2 by 1.8% pts while Test Set R2 stats decreased by 1.4% pts.

To understand the change on the important predictor variables, lets look at the below table:

	Ridge Regression		**Lasso Regression**	
Rank	Optimal Alpha (α = 6.59)	Double Alpha (α = 13.18)	Optimal Alpha (α = 0.0003)	Double Alpha (α = 0.0006)
1	**OverallQual**	**OverallQual**	**GrLivArea**	**GrLivArea**
2	**2ndFlrSF**	**Neighborhood_NoRidge**	**OverallQual**	**OverallQual**
3	**GrLivArea**	**GrLivArea**	**Neighborhood_NoRidge**	**Neighborhood_NoRidge**
4	**Neighborhood_NoRidge**	**2ndFlrSF**	**GarageCars**	**GarageCars**
5	**RoofMatl_WdShngl**	**1stFlrSF**	**RoofMatl_WdShngl**	**Neighborhood_NridgHt**
6	1stFlrSF	TotRmsAbvGrd	Neighborhood_NridgHt	BsmtExposure_Gd
7	GarageCars	FullBath	BsmtExposure_Gd	Fireplaces
8	FullBath	GarageCars	BsmtFullBath	BsmtFullBath
9	TotRmsAbvGrd	RoofMatl_WdShngl	Neighborhood_Crawfor	Neighborhood_Crawfor
10	Neighborhood_NridgHt	Fireplace	Neighborhood_Somerst	Neighborhood_Somers

The observations are as follows:

- For both Ridge and Lasso, we see the top 5 predictors remain more or less the same even when alpha is doubled. There is occasional shift of placed within top 5 predictors.
- For the top 10 predictors, in Lasso Regression, we see the same variables in the two groups (model with optimal alpha & model with double alpha).
- However in Ridge, one variable "Fireplace" has come in top 10 predictor, replacing "Neighborhood_NridgHt" from the model that has the optimal alpha.

- Overall, no significant difference observed in predictor variables for both ridge and lasso, when the optimal alphas were doubled.

When alpha has been doubled, the top 5 predictor variables are highlighted in blue below:

	Ridge Regression		Lasso Regression	
Rank	Optimal Alpha (α = 6.59)	Double Alpha (α = 13.18)	Optimal Alpha (α = 0.0003)	Double Alpha (α = 0.0006)
1	OverallQual	**OverallQual**	GrLivArea	**GrLivArea**
2	2ndFlrSF	**Neighborhood_NoRidge**	OverallQual	**OverallQual**
3	GrLivArea	**GrLivArea**	Neighborhood_NoRidge	**Neighborhood_NoRidge**
4	Neighborhood_NoRidge	**2ndFlrSF**	GarageCars	**GarageCars**
5	RoofMatl_WdShngl	**1stFlrSF**	RoofMatl_WdShngl	**Neighborhood_NridgHt**

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

To determine, which to choose between Ridge and Lasso, we will compare their train and test R2 metrics.

Metrics	Linear Regression	Linear Regression (RFE)	Ridge	Lasso
Train R2	9.49E-01	0.907513	0.889554	0.877653
Test R2	-2.69E+21	0.796727	0.866821	0.857589
RSS	6.27E-01	1.138057	1.359045	22686.1208
MSE	6.14E-04	0.001115	0.001331	0.001475
RMSE	2.48E-02	0.024772	0.036484	0.038399

From above table, we can conclude that **Ridge is a slightly better model than Lasso** since Training R2 is **~89% and Test R2 is ~87%,** both of which are higher than Lasso model and also the difference between them is only 2% pts.

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

The 5 most important predictor variables in Lasso model (original) were - 'GrLivArea','OverallQual', 'Neighborhood_NoRidge', 'GarageCars', 'RoofMatl_WdShngl'.

Removing these variables and re running the Lasso Model with the same alpha = 0.0003, we get the new set of top 5 predictors. They are as follows:

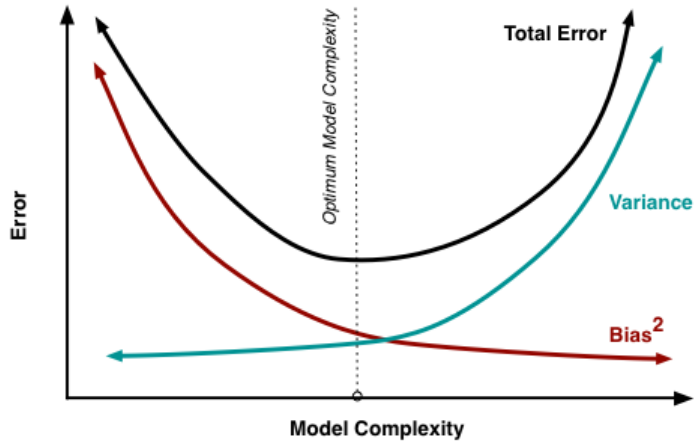
Rank	Old Top Predictor	New Top Predictor
1	GrLivArea	**1stFlrSF**
2	OverallQual	**2ndFlrSF**
3	Neighborhood_NoRidge	**GarageArea**
4	GarageCars	**MasVnrArea**
5	RoofMatl_WdShngl'	**BsmtExposure_Gd**

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 has the following characteristics:



- Low Bias and Low Variance, which means an optimal model with optimum no. of features so as to have the minimal total error. This leads to Optimum Model Complexity
- Such type of models don't have polynomial features, which can lead to overfitting
- To achieve the right amount of model complexity, there are regularization techniques such as Ridge and Lasso Regression, which reduce the betas (lasso reduces them to zero), and thereby reduce overfitting of the training dataset

Implications of such model

- The R2 on training dataset will reduce. However, this is on purpose so as to reduce overfitting.
- The Betas of the predictors will reduce to enable reduction in R2.
- The Test Set R2 will increase, now that model is better able to predict the test dataset.
- The difference in R2 for Training set and Test Set will decrease, leading to increase in model confidence for predicting values for unseen data.
- **Because of the points mentioned above, the accuracy of the model increases post regularization.**