Advanced Regression – Assignment **By – Sankalp Gupta**

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:

(a) The raw data contained 81 columns. Of these, SalePrice is the Target column and the remaining 80 are features. Post EDA and Data cleaning, the number of features became 234.

(b) After performing a 5-fold cross-validation grid-search, the best Regularization parameter, lambda is:

Ridge Regression: 5.3

Lasso Regression: 59.0

(c) The above search was done for lambda ranging from 0.00001 to 1000. There were a total of 29 possible candidates.

(d) The most important predictor variables (top 10) are:

S#	Ridge (Alpha = 5.3)	Ridge (Alpha = 10.6)	Lasso (Alpha = 59.0)	Lasso (Alpha = 118.0)
1	OverallQual	OverallQual	GrLivArea	GrLivArea
2	WdShngl_RoofMatl	StoneBr_Nbrhood	WdShngl_RoofMatl	WdShngl_RoofMatl
3	StoneBr_Nbrhood	TotRmsAbvGrd	OverallQual	OverallQual
4	TotRmsAbvGrd	GarageCars	StoneBr_Nbrhood	StoneBr_Nbrhood
5	GrLivArea	WdShngl_RoofMatl	LotArea	GarageCars
6	GarageCars	GrLivArea	BsmtFinSF1	NoRidge_Nbrhood
7	2ndFlrSF	FullBath	OverallCond	NridgHt_Nbrhood
8	FullBath	2ndFlrSF	GarageCars	LotArea
9	NoRidge_Nbrhood	NoRidge_Nbrhood	NoRidge_Nbrhood	OverallCond
10	1stFlrSF	NridgHt_Nbrhood	TotRmsAbvGrd	TotRmsAbvGrd

Notes on the above table:

1. 7 of the top 10 features are common between Lasso and Ridge Models. This means that the models are stable and not random.
2. Changing Alpha changed the order of features, which means the relative coefficient values have changed. However, the features are more or less the same.

3. The above model was arrived at by using all the features and all the property data. There were some features which had mostly a single value in the column, and there were outliers in the target variable. Regression was done after removing these single valued columns, as well as outliers in the target variable. This was done to check the stability of the model. Here are the results of the same:

Case 1: All Features are used. However, outliers are removed from target variable

S#	Ridge (Alpha = 5.3)	Lasso (Alpha = 59.0)
1	OverallQual	GrLivArea
2	TotRmsAbvGrd	OverallQual
3	GarageCars	OverallCond
4	GrLivArea	GarageArea
5	2ndFlrSF	TotalBsmtSF
6	OverallCond	NoRidge_Nbrhood
7	GarageArea	TotRmsAbvGrd
8	NoRidge_Nbrhood	NridgHt_Nbrhood
9	1stFlrSF	LotArea
10	NridgHt_Nbrhood	GarageCars
	Note: 7 of the earlier features were retained	Note: 7 of the earlier features were retained

Case 2: Single Value Features are removed. However, all house rows are retained

S#	Ridge (Alpha = 5.3)	Lasso (Alpha = 59.0)
1	OverallQual	GrLivArea
2	StoneBr_Nbrhood	LotArea
3	GarageCars	OverallQual
4	GrLivArea	StoneBr_Nbrhood
5	TotRmsAbvGrd	GarageCars
6	2ndFlrSF	NoRidge_Nbrhood
7	FullBath	NridgHt_Nbrhood
8	NoRidge_Nbrhood	TotRmsAbvGrd
9	1stFlrSF	OverallCond
10	NridgHt_Nbrhood	FullBath
	Note: 9 of the earlier features were retained	Note: 8 of the earlier features were retained

Case 3: Single Value Features are removed. Also, outliers are removed

S#	Ridge (Alpha = 5.3)	Lasso (Alpha = 59.0)
1	OverallQual	GrLivArea
2	TotRmsAbvGrd	OverallQual
3	GarageCars	GarageCars
4	2ndFlrSF	TotRmsAbvGrd
5	GrLivArea	OverallCond
6	OverallCond	NoRidge_Nbrhood
7	NoRidge_Nbrhood	NridgHt_Nbrhood
8	GarageArea	Ex_BsmtQual
9	NridgHt_Nbrhood	2ndFlrSF
10	1stFlrSF	Ext_BrkFace
	Note: 7 of the earlier features were retained	Note: 6 of the earlier features were retained

Based on the regression using different strategies, we find that the top 10 features are almost similar and hence, we can say that the model is quite stable.

Residuals analysis was also done. All the assumptions were found to be true.

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:

The decision of which regression model to choose will depend on the stability of the model and a study of the business domain. This will involve the following steps:

1. Compare the top 10 features of each model. Confirm that most of the them are same. In this case, we notice that 7 out of the top 10 features are same. This gives us confidence that our model is stable
2. Check with a business domain expert on whether the top 7 common features are indeed the main decision making factors.
3. Of the remaining 3 factors, check with the business domain expert, as to which ones are more important.
4. Check the coefficient values of the remaining 3 features.
5. Do another round of study from step 1, with top 15 features instead of top 10.
6. Finally, based on the above knowledge, and based on the domain expert's advise, decide which of the models will be better.
7. Once the number of top features and the model is decided, re-model to get the fresh value of coefficients. This model can now be used to make predictions.

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 top 5 features for both Ridge and Lasso, after removing the original Top 5 predictors, are given in the table below:

S#	Ridge	Lasso
1	1stFlrSF	1stFlrSF
2	2ndFlrSF	2ndFlrSF
3	GarageCars	TotalBsmtSF
4	FullBath	BsmtFinSF1
5	NoRidge_Nbrhood	GarageCars

Note:

1. The next 5 features in the previous Ridge model have now become the top 5

features, however, their order has changed, which means their relative coefficient values have changed

2. However, in the case of Lasso, only 2 of the original next 5 features are there. The remaining 3 are totally new.

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) To confirm whether the model is robust and generalisable, the following three points should be studied:

1. Different models should produce similar values for coefficients and top 10-15 features.
2. The comparison of R2 score of Training vs Test data should be close to each other. R2 score of Test data should not fall too much from the R2 score of the Training Data.
3. The top 10-15 features, as indicated by the model, should agree with the intuitive knowledge of the business domain expert.

(b) A robust model can be used to make predictions that are reliable and can be used to make (i) repeatable business decisions, and (ii) within an acceptable margin of error. If the model is not robust then the predictions made by the model will result in business decisions that are not informed decisions, but are like the throw of a dice.

~ END of Document ~