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 230.

(b) A study of multi-collinearity of the top 25 features revealed many highly collinear features. 6 of these were removed before beginning with regression.

(c) There were many single-valued columns, or almost single-valued columns. A total of 12 such features were removed

(d) There were 56 outliers in Sales Price. The same were removed too

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

Ridge Regression: 4.0

Lasso Regression: 0.0001

(f) The above search was done for lambda ranging from 0.00001 to 1000.

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

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

S#	Ridge (Alpha = 4.0)	Lasso (Alpha = 0.0001)	Ridge (Alpha = 8.0)	Lasso (Alpha = 0.0002)	Collinear
1	GarageCars	1stFlrSF	GarageCars	1stFlrSF	GarageCars
2	2ndFlrSF	2ndFlrSF	TotRmsAbvGrd	2ndFlrSF	FullBath
3	TotRmsAbvGrd	GarageCars	2ndFlrSF	GarageCars	TotalBsmtSF
4	1stFlrSF	OverallCond	1stFlrSF	OverallCond	1stFlrSF
5	OverallCond	NoRidge_Nbrhood	OverallCond	TotRmsAbvGrd	PConc_Foundation
6	NoRidge_Nbrhood	TotRmsAbvGrd	NoRidge_Nbrhood	NoRidge_Nbrhood	Gd_KitchenQual
7	FullBath	WdShngl_RoofMatl	FullBath	StoneBr_Nbrhood	TotRmsAbvGrd
8	NridgHt_Nbrhood	StoneBr_Nbrhood	NridgHt_Nbrhood	NridgHt_Nbrhood	Fireplaces
9	StoneBr_Nbrhood	NridgHt_Nbrhood	Ex_BsmtQual	Ex_BsmtQual	SubClass_60
10	Ex_BsmtQual	FV_Zone	StoneBr_Nbrhood	Somerst_Nbrhood	GLQ_BsmtFinType1

From the table above, we note:

1. 8 of the top 10 features are common between Lasso and Ridge Models. This indicates 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 models were arrived at by using all the features but removing outliers in Sales Price data. There were some features which had mostly a single value in the column. They were removed before regression.
4. 4 of the top 10 features have high collinearity with Sales Price. But the others 6 features with high collinearity, did not appear in either Ridge or Lasso model.
5. The Alpha value for Ridge and Lasso is very different from each other. This indicates that the two models are quite independent of each other.
6. Both the models were able to explain about 86% variance in the Sales Price, with a Root Mean Square Error of about 7%.

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 8 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 8 common features are indeed the main decision making factors.
3. Of the remaining 4 features – 2 each in Ridge and Lasso, check with the business domain expert, as to which ones are more important.
4. Check the coefficient values of the remaining 6 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	FullBath	WdShake_RoofMatl
2	NoRidge_Nbrhood	WdShngl_RoofMatl
3	BedroomAbvGr	CompShg_RoofMatl
4	TotalBsmtSF	LotArea
5	NridgHt_Nbrhood	BsmtFinSF1

Note:

1. In the Ridge model: Out of the next 5 features in the previous Ridge model, 3 have now come in the top 5 features, however, their order has changed, which means their relative coefficient values have changed. 2 new features have appeared in the top 5, replacing 2 of the earlier top 10 features
2. In the Lasso model, only 1 of the original features remained in the top 5. 4 are new. Of the new features, none is common to Lasso and Ridge.

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 ~