Assignment: Advanced Regression

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?

Ans:

- Alpha for ridge regression: 10
- Alpha for lasso regression: 0.001

After doubling the value of alpha for both ridge and lasso regressions, the metrics have undergone slight changes, as shown in below tables-

Metric	Ridge regression		Lasso regression	
	Alpha = 10	**Alpha =20**	**Alpha = 0.001**	**Alpha =0.002**
R2 Score (Train)	0.944	0.932	0.914	0.892
R2 Score (Test)	0.893	0.893	0.900	0.888
RSS (Train)	4.722	5.563	7.038	8.829
RSS (Test)	3.886	3.894	3.623	4.065
MSE (Train)	0.069	0.075	0.085	0.095
MSE (Test)	0.096	0.096	0.093	0.098

- R2 score (train) has decreased for both ridge and lasso regression models with change in alpha
- RSS(train) has increased for both ridge and lasso regression models with doubling of the alpha value
- The effect of doubling of alpha is however not significant on the test dataset

With change in alpha, there is a noticeable change in the top features. Following tables highlight the comparison of top 15 variables with lasso and ridge regressions. With change in alpha, the hierarchy of the features have changed. The features are grouped according to descending order, with highest magnitude of coefficient at the top.

Sl.no	Lasso regression	
	Alpha= 0.001	**Alpha=0.002**
1	GrLivArea	GrLivArea
2	TotalBsmtSF	TotalBsmtSF
3	Neighborhood_Crawfor	GarageArea
4	FullBath	LotArea
5	OverallQual_Very Good	GarageCars
6	OverallQual_Excellent	Neighborhood_Crawfor
7	LotArea	OverallQual_Very Good
8	GarageArea	FullBath
9	GarageCars	BsmtFinSF1
10	HalfBath	Functional_Typ

11	Neighborhood_Somerst	HalfBath
12	Functional_Typ	BsmtFinType1_GLQ
13	OverallQual_Good	OverallQual_Good
14	BsmtCond_Gd	GarageType_Attchd
15	BsmtFinSF1	OverallCond_Good

Sl.no	**Ridge regression**	
	Alpha= 10	**Alpha=20**
1	GrLivArea	GrLivArea
2	TotalBsmtSF	TotalBsmtSF
3	1stFlrSF	1stFlrSF
4	FullBath	FullBath
5	Neighborhood_Crawfor	Neighborhood_Crawfor
6	OverallQual_Excellent	GarageArea
7	OverallQual_Very Good	LotArea
8	LotArea	OverallQual_Very Good
9	GarageArea	OverallQual_Excellent
10	HalfBath	GarageCars
11	BsmtCond_Gd	HalfBath
12	SaleCondition_Normal	SaleCondition_Normal
13	GarageCars	2ndFlrSF
14	Neighborhood_ClearCr	TotRmsAbvGrd
15	Exterior1st_BrkFace	Exterior1st_BrkFace

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?

Ans: It is preferable to use the lasso regression model, as apart from the metrics being nearly the same as that of ridge regression model, there is feature selection as well. Having fewer variables would help in better business understanding. However even with lasso regression, there are considerable number of variables which are correlated and therefore it might be necessary to further filter the features using RFE and manual removal of features based on VIF values.

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?

Ans: Following are the 5 most important predictors, with their coefficients:

Predictors	**Coefficient**
TotRmsAbvGrd	0.141351
BsmtFinSF1	0.111579
OverallQual_Excellent	0.105638

GarageArea	0.103947
LotArea	0.101561

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?

Ans: To make a model robust and generalisable, following points can be considered-

- Outlier removal or feature engineering: this involves a good investigation of the raw data, to check for outliers, missing data or incorrect data. The machine learning model coefficients might end up being incorrect and very sensitive to changes in training data, if this aspect is not addressed
- Remove multicollinearity: highly correlated data need to be removed. Linear regression model in this problem has overfit the data and the test dataset metrics were not good, this is probably due to some independent variables being highly correlated. The same was to some extent solved through ridge or lasso regression models
- Fewer variables in the model: having fewer variables in the model can help in better interpretation of the model, however this might come at the cost of accuracy, since some variables which may not have a significant impact get discarded
- Bias-variance tradeoff has to be managed to get a model that is not too complex (high variance, overfitting etc…) and accurate as well