Assignment-based Questions

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. The optimal values for Alpha are:

- Ridge – 6.8
- Lasso – 0.0004

If we choose double the values then for

Ridge - The penalty will be more and the coefficient will reduce but the values will not be 0 it will be tending towards zero

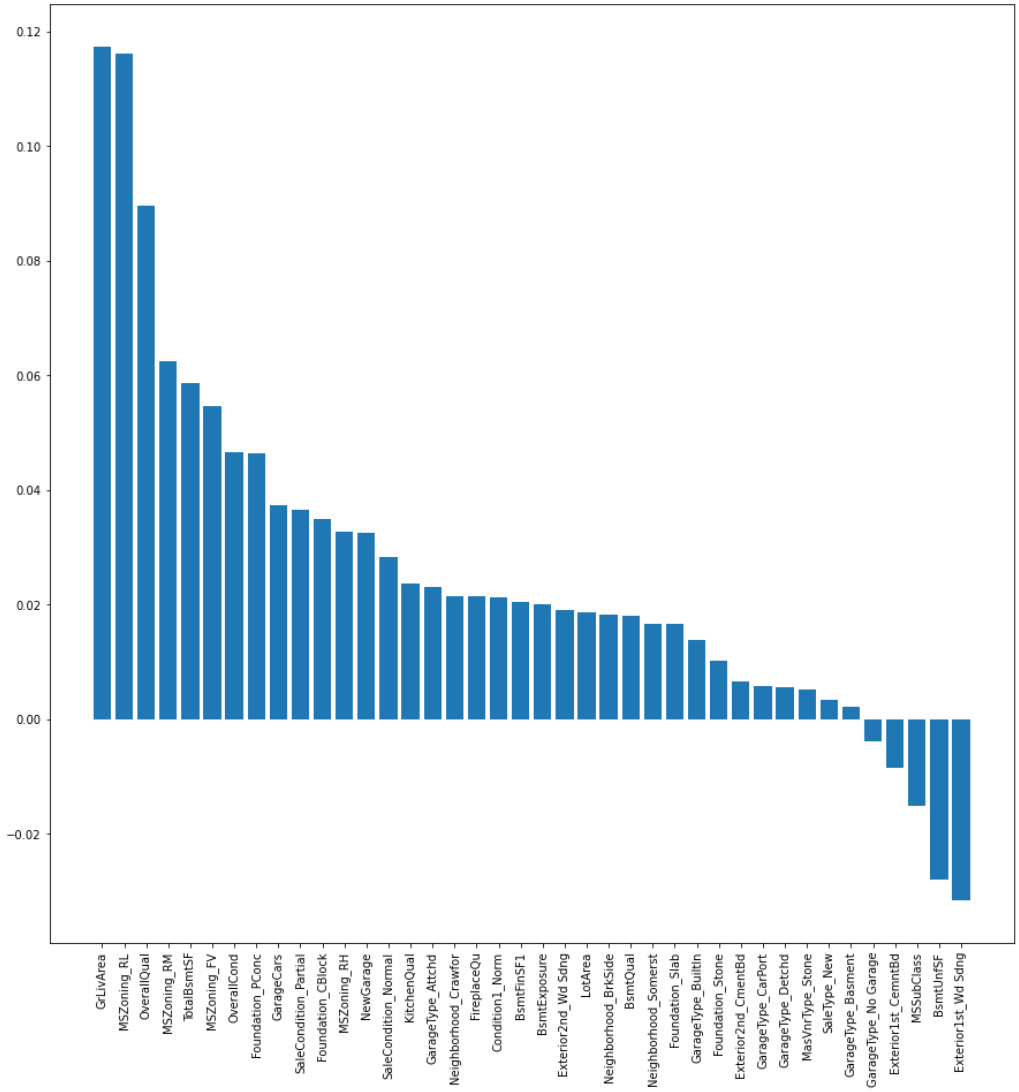
$$\widehat{\beta_R} = \sum (y_{true} - y_{pred})^2 + \lambda \cdot \sum_{i=1}^{n} \beta_i^2$$

Lasso – The penalty will be high and unlike ridge here few more coefficients will become nearly zero or zero

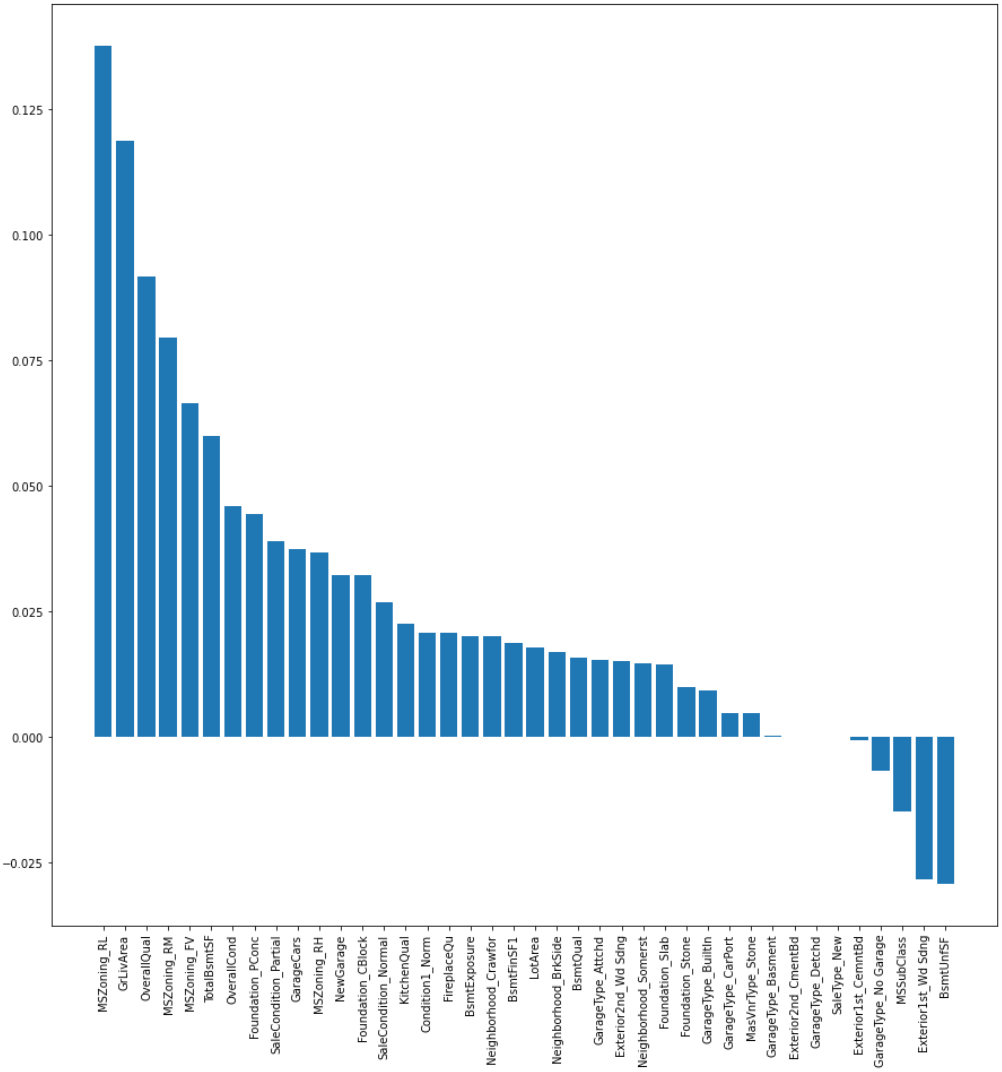
$$\widehat{\beta_L} = \sum (y_{true} - y_{pred})^2 + \lambda \cdot \sum_{i=1}^{n} |\beta_i|$$

The most significant variables after the change is :

- Ridge


0.12
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
GrLivArea
MSZoning_RL
OverallQual
MSZoning_RM
TotalBsmtSF
MSZoning_FV
OverallCond
Foundation_PConc
GarageCars
SaleCondition_Partial
Foundation_CBlock
MSZoning_RH
NewGarage
SaleCondition_Normal
KitchenQual
GarageType_Attchd
Neighborhood_Crawfor
FireplaceQu
Condition1_Norm
BsmtFinSF1
BsmtExposure
Exterior2nd_Wd Sdng
LotArea
Neighborhood_BrkSide
BsmtQual
Neighborhood_Somerst
Foundation_Slab
GarageType_BuiltIn
Foundation_Stone
Exterior2nd_CmentBd
GarageType_CarPort
GarageType_Detchd
MasVnrType_Stone
SaleType_New
GarageType_Basment
GarageType_No Garage
Exterior1st_CemntBd
MSSubClass
BsmtUnfSF
Exterior1st_Wd Sdng

- Lasso


0.125
0.100
0.075
0.050
0.025
0.000
−0.025
MSZoning_RL
GrLivArea
OverallQual
MSZoning_RM
MSZoning_FV
TotalBsmtSF
OverallCond
Foundation_PConc
SaleCondition_Partial
GarageCars
MSZoning_RH
NewGarage
Foundation_CBlock
SaleCondition_Normal
KitchenQual
Condition1_Norm
FireplaceQu
BsmtExposure
Neighborhood_Crawfor
BsmtFinSF1
LotArea
Neighborhood_BrkSide
BsmtQual
GarageType_Attchd
Exterior2nd_Wd Sdng
Neighborhood_Somerst
Foundation_Slab
Foundation_Stone
GarageType_BuiltIn
GarageType_CarPort
MasVnrType_Stone
GarageType_Basment
Exterior2nd_CmentBd
GarageType_Detchd
SaleType_New
Exterior1st_CemntBd
GarageType_No Garage
MSSubClass
Exterior1st_Wd Sdng
BsmtUnfSF

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. We will be choosing the Lasso regression model. The below image depicts the evaluation parameters for each models - linear regression, ridge regression & lasso regression. The R2 score along with the RSS and RMSE we see Ridge and Lasso has an edge over normal regression. In between ridge and lasso we chose lasso because it does feature selection and it did here too, reducing coefficients of 3 features to zero.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.919620	0.918796	0.919029
1	R2 Score (Test)	0.900453	0.904719	0.903863
2	RSS (Train)	10.829300	10.940409	10.909017
3	RSS (Test)	5.887702	5.635375	5.686055
4	MSE (Train)	0.011533	0.011651	0.011618
5	MSE (Test)	0.014610	0.013984	0.014109
6	RMSE (Train)	0.107391	0.107940	0.107785
7	RMSE (Test)	0.120871	0.118252	0.118783

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. Old Top 5 :

	Features	Lasso Coeff
16	MSZoning_RL	0.164657
9	GrLivArea	0.119434
17	MSZoning_RM	0.104387
2	OverallQual	0.090025
14	MSZoning_FV	0.080678

New Top 5 :

	Features	Lasso Coeff
0	TotalBsmtSF	0.112081
5	GarageCars	0.093414
12	FireplaceQu	0.066460
2	OverallCond	0.058453
9	KitchenQual	0.056176

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. The following can help achieve it :

- A model needs to be effective enough so as to not get impacted by outliers. It should be made robust and generalizable by using proper treatments against it. We must carry out proper analysis of outliers and decide how to remove them may be using standard deviation intervals or if some transformations can be used to handle them like Log, etc.

- Regularization is also a key to better modelling. It generalizes the overall model to learn in an effective manner so that it does not overfit on the data and perform good on unseen data. It also handles multicollinearity among the features. Too much weightage on such variables can have huge impact on overall model.

If these issues are not handled in a proper way it has a direct impact on accuracy like outliers can lead to increase in error variance and poor statistical test results. If the model is not generalized enough it can lead to overfitting thus poor test results and the model will not be robust enough to perform on unseen data.