Assignment 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?

The optimal value of alpha is:

Ridge – 2 Lasso – 0.0003

Doubling the value of alpha: Ridge = 4 and Lasso = 0.006

	Ridge		Lasso	
Alpha	2	4	0.0003	0.0006
Training R2	0.9556	0.9512	0.9509	0.9444
Testing R2	0.8629	0.8647	0.8584	0.8602

With change in alpha there is a small change in R2 of training and testing, as alpha increases we observe slight decrease in training R2 and increase in testing R2.

The most important predictor variables after doubling alphas are: Overall Quality, Living area, First floor square feet, Overall condition and basement size.

	Ridge Co-Efficient
GrLivArea	0.315524
OverallQual	0.301540
1stFlrSF	0.297083
OverallCond	0.215423
2ndFlrSF	0.209172
LotArea	0.197095
TotalBsmtSF	0.194189
BsmtFinSF1	0.184116
MSZoning_FV	0.127374
GarageArea	0.116371
Neighborhood_StoneBr	0.112973

	Ridge Doubled Alpha Co-Efficient
OverallQual	0.277848
GrLivArea	0.266122
1stFlrSF	0.253249
OverallCond	0.187328
TotalBsmtSF	0.181465
2ndFlrSF	0.173522
BsmtFinSF1	0.170865
LotArea	0.146173
GarageArea	0.114531
FullBath	0.109413

	Lasso Co-Efficient
GrLivArea	0.985011
OverallQual	0.407253
TotalBsmtSF	0.284632
LotArea	0.249869
OverallCond	0.243774
BsmtFinSF1	0.138570
ScreenPorch	0.103165
Neighborhood_Crawfor	0.102674
GarageArea	0.093500
Neighborhood_StoneBr	0.090528

	Lasso Doubled Alpha Co-Efficient
GrLivArea	0.983680
OverallQual	0.456306
TotalBsmtSF	0.295024
OverallCond	0.227212
LotArea	0.200271
BsmtFinSF1	0.136787
Neighborhood_Crawfor	0.097478
GarageArea	0.087142
ScreenPorch	0.079105
GarageCars	0.076635

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?

The optimal value of alpha is:

- Ridge: 2
- Lasso: 0.0003

Observing the R2 metrics for both Ridge and Lasso are around same 95% for training and 86% for testing. Since there is no much difference in metrics, I would choose Lasso as it will eliminate the features as well making the model more robust.

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?

The five most important predictor variables in Lasso currently are: Living area, Overall quality, TotalvBasement squarefoot, Lot Area and Overall condition of the property.

	Lasso Co-Efficient
GrLivArea	0.985011
OverallQual	0.407253
TotalBsmtSF	0.284632
LotArea	0.249869
OverallCond	0.243774

Removing these columns from the training set and running the model again to see the next 5 most important predictor variables.

The most important predictor variables are as follows: first floor square foot, second floor squ are foot, Basement square foot, Garage Area and Screen Proch.

	Lasso Co-Efficient
1stFlrSF	0.856653
2ndFlrSF	0.541946
BsmtFinSF1	0.253769
GarageArea	0.120873
ScreenPorch	0.119514

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?

The model complexity increases or say depends on Magnitude of the coefficients and number of coefficients. Although more data can give us more information, it makes the model more complex and if the data is not fairly collected may cause bias and overfitting on training data.

Few steps we can take to make model more robust are:

- Adding more relevant data.
- Feature Selection: check which features are more significant than others in prediction and keep those and remove highly correlated features.
- Treat missing data and outliers: Data may be missing because that feature may not be present for that. We will have to understand business and then decide on how to treat missing data whether it has to be imputed with mean/median/mode or mention absence of that feature which may affect prediction.
  Outliers have to be treated as model coefficients may be sensitive and help improve accuracy of the model.
- Feature Engineering: Features may need to be transformed to make it more linear or to correct the skewness of the data. Also combining few features into one to make simple feature helping business understand better will make our model more robust.
- Visualization: It helps us understand the features better and how a particular feature is impacting output variable and how are the features correlated.
- Model Selection:

Model coefficients that we obtain from ordinary least square can be unreliable when only few of the predictor variable of the model is significantly related to response variable.

During that time, we use Regularization, where we add penalty to the model's cost function.

Regularization helps with managing model complexity by shrinking the model coefficients towards zero or minimizing it. Thus, it avoids the model to be more complex and reduces overfitting.

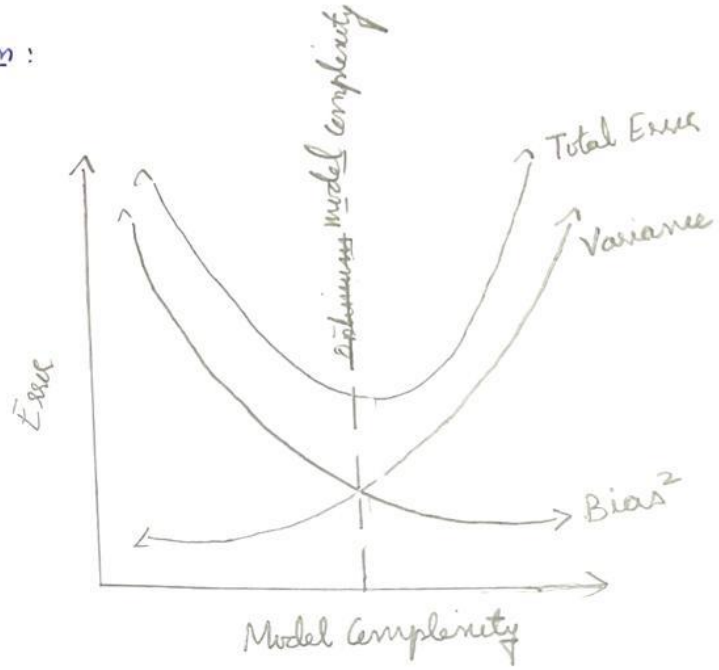
We have to have trade off between bias and variance.

To minimize the variance little bias has to be made to get a more generic and robust model.

There are different types of regularization which we can apply, the most common ones are Ridge and Lasso.

The below figure shows the optimal model complexity to be maintained by trading off bias and variance. The formula gives how penalty is added to cost function in Ridge method.

Regularization :



Cost function of OLS :

$$RSS : \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Cost function for Ridge:

Shrinkage Penalty

$$Cost : \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

β_j^2 => sum of squared model coefficient

λ => Tuning Parameter

if $\lambda = 0$ (no penalty) => result in overfitting