Subjective Questions – 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?

ANSWER:

- With the cross validation approach, I got optimal alpha for ridge as **22** and **0.0021** for Lasso.
- When we double the value of alpha for our **Ridge Regression**, the model will apply more penalty on the curve and try to make the model more generalized that is making model simpler and not suffer from overfitting. Error decrease in training score and testing result is optimal.

- The most important predictor variables after the changes is implemented for Ridge Regression:
  1. SaleCondition_Normal
  2. LotArea
  3. Condition1_Norm
  4. YearBuilt
  5. OverallGrade
  6. Neighborhood_Crawfor
  7. CentralAir_Y
  8. Neighborhood_StoneBr
  9. GrLivArea-Sq
  10. AllSF-Sq

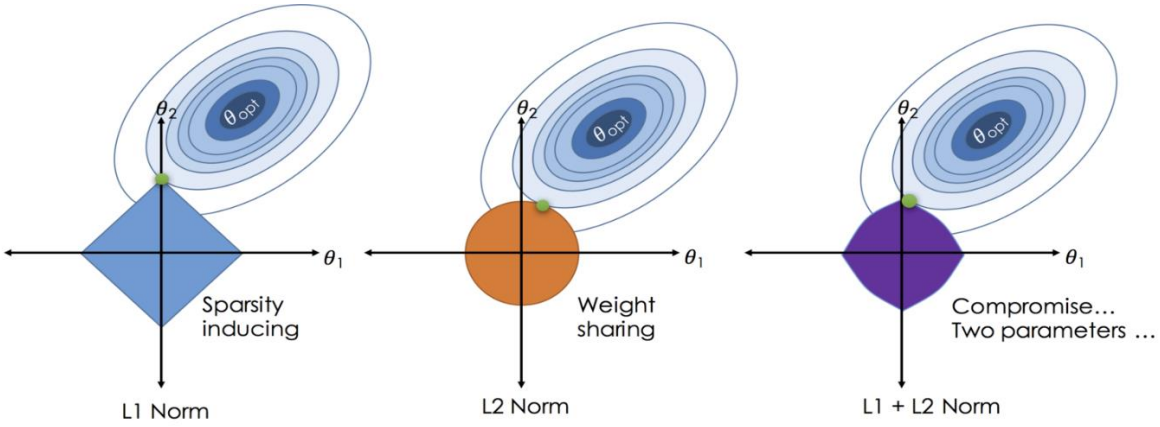
- Similarly, when we increase the value of alpha for **Lasso Regression** we try to penalize more and more coefficient of the variable will have reduced to zero, when we increase the value, R2 square also decreases. Additionally, 23 variables dropped by doubling alpha in lasso.

- The most important predictor variables after the changes is implemented for Lasso Regression:
  1. AllSF-Sq
  2. SaleCondition_Normal
  3. YearBuilt
  4. OverallGrade
  5. LotArea
  6. Condition1_Norm
  7. GrLivArea-Sq
  8. OverallQual
  9. BoughtOffPlan
  10. Neighborhood_Crawfor

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:

I will prefer Lasso over Ridge, because Lasso helps in variable selection. Simpler model is always better than complex models. It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretable. With this I prefer Lasso Regression.



Ridge regression (L2-Regularization): Uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values get penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes most of all the variables in final model unlike Lasso Regression.

Lasso regression (L1-Regularization): Uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does **variable selection**. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

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?

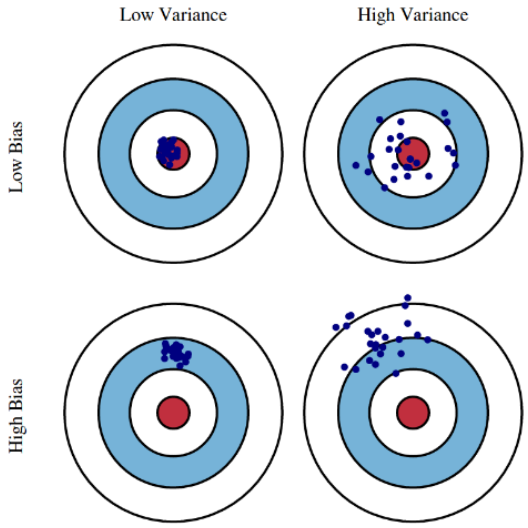
ANSWER:

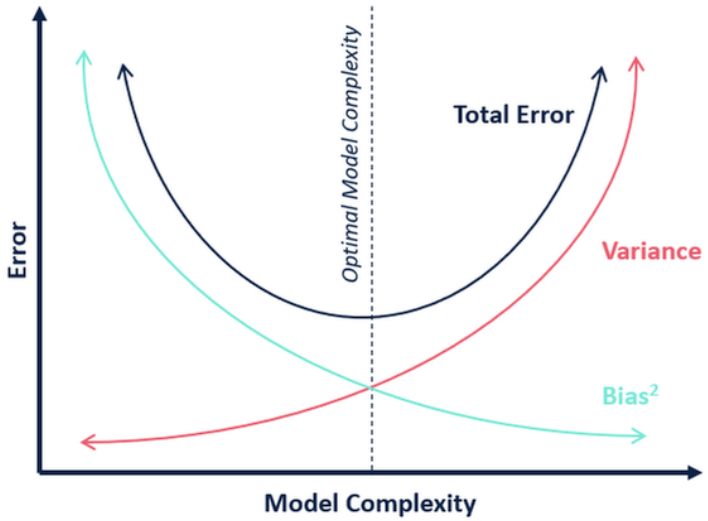
Five most important predictor variables are:

1. Condition1 - Proximity to various conditions(Normal)
2. YearBuilt - Original construction date
3. LotArea - Lot size in square feet
4. OverallGrade – Overall quality & condition of the property
5. CentralAir – Presence of central air conditioning

4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

ANSWER:





Simpler models are always better than complex models, though accuracy will drop but it will be more robust and generalizable. It can be also understood using bias-variance trade-off. The simpler the model more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and validation datasets. The accuracy is more or less same in training and testing data.

Bias: Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Variance: Variance is the amount that the estimate of the target function will change given different training data.

It is important to balance between Bias and Variance to avoid overfitting and Under fitting of data.