House Price Prediction Assignment Subjective Questions and Answers

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?

Ans-

1. Optimal Value of alpha for ridge and lasso Regression
    - For Ridge regression :1.0
    - For Lasso Regression :0.0001
2. R2Score on training data has decreased but it has increased in testing data.
3. For Lasso, on doubling the alpha value, the test set R-Squared goes down and for Ridge, on doubling the alpha value the test set R-Squared goes down.

Here is a slight change in the important predictor variables in case of Lasso, but there are noticable differences in case of Ridge in terms of important predictor variables.

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?

Ans.

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably.

Ridge regression, 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 gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, 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.

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?

Ans.

Now the top 5 variables are:

1. LotFrontage, 2) Overall Condition, 3) MSZoning_RH , 4)Overall quality,5) Garage Area

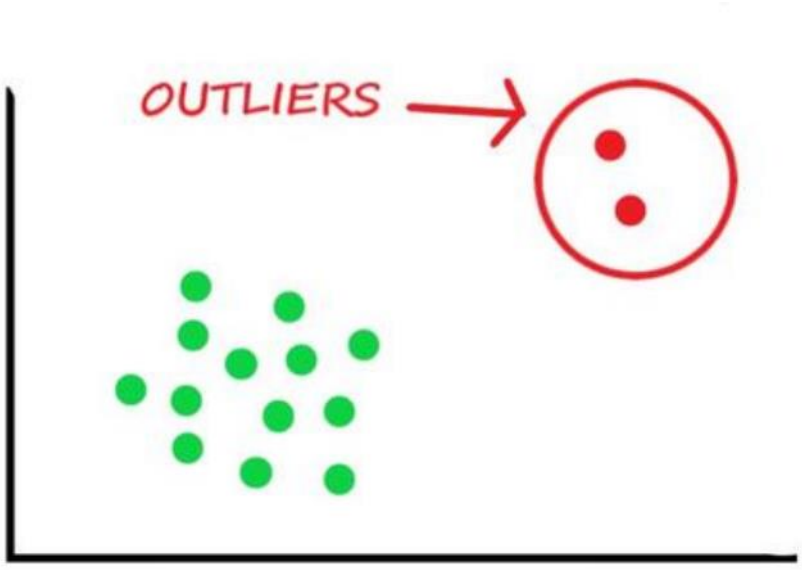
   By excluding top predicted variables, automatically r2 square will be dropped and will reduce model predictive power

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?

Ans.

When we say a model is robust, it means that it needs to be immune to outliers. When we train on data that has outliers, it will skew the results of the model. It can potentially harm the predictive power of the model. Outliers are the points that are distanced from other observations.



Now, when we try to fit a best fit line through the points, the coefficients will be much smaller in case of a regularized regression as it will try to penalize for the high error. This can make the coefficients of the

model unreliable hence making the predictions on unseen data inaccurate. This might make someone think that deleting outliers would be the right strategy, but we must treat them carefully instead. Now, Outliers can be misread observations or can be intended. We can use several methods to detect outliers like using a Box plot or a Scatter plot. After detecting we can use any suitable methods like Inter Quartile Range or Z-Score (Standard Deviation) method to remove outliers. We can even replace the outliers with a more suitable 99th percentile or 25th percentile value. This will help in increasing the prediction power of the model on unseen data as well. Hence making it more generalizable. We can also use regularization techniques of Ridge and Lasso that penalize model for overfitting the training data also minimizing the effects of outliers.