PROBLEM STATEMENT - 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?

Answer:

Regularizing coefficients is an important aspect of improving prediction accuracy while reducing variance and maintaining model interpretability. Ridge regression employs a tuning parameter called lambda to apply a penalty that is proportional to the square of the magnitude of the coefficients. This lambda value is typically determined through cross-validation. The penalty term helps minimize the residual sum of squares, aiming to keep it small. By multiplying lambda with the sum of squares of the coefficients, Ridge regression penalizes coefficients with larger values more heavily. As lambda increases, the variance in the model decreases while the bias remains relatively constant. Unlike Lasso regression, Ridge regression includes all variables in the final model.

On the other hand, Lasso regression also uses a tuning parameter lambda as a penalty. However, in Lasso regression, the penalty is the absolute value of the magnitude of the coefficients. As lambda increases, Lasso regression progressively shrinks the coefficients towards zero. One notable characteristic of Lasso regression is its variable selection capability. When lambda is small, Lasso regression behaves similarly to simple linear regression. However, as the lambda value increases, shrinkage occurs, causing some coefficients to become exactly zero. Consequently, Lasso regression neglects variables that have coefficients with a value of zero, thus performing variable selection in addition to regularization.

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?

Answer:

Choosing Lasso regression as the preferred method is a valid decision. Lasso regression not only provides regularization to improve model accuracy and generalization but also offers the advantage of feature selection. By setting appropriate lambda values, Lasso regression effectively removes unwanted or irrelevant features from the model.

The feature selection capability of Lasso regression is valuable because it simplifies the model by discarding unnecessary variables. This not only enhances the interpretability of the model but also reduces the risk of overfitting. By eliminating irrelevant features, Lasso regression focuses on the most important predictors, leading to a more concise and efficient model.

It is worth noting that while Lasso regression can remove irrelevant features without significantly impacting model accuracy, the specific impact on model performance depends on the dataset and the lambda value chosen. It is recommended to fine-tune the lambda value through techniques such as cross-validation to strike the right balance between feature selection and prediction accuracy.

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?

Answer:

Those 5 most important predictor variables that will be excluded are :-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageAre

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?

Answer:

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the 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 test data i.e. the accuracy does not change much for training and test data

Bias: Bias refers to the error in the model when it is unable to capture the underlying patterns in the data. A high bias indicates that the model is too simplistic and struggles to learn the intricate details in the data. As a result, the model performs poorly on both the training and testing data, leading to low accuracy.

Variance: Variance, on the other hand, refers to the error in the model when it becomes overly complex and starts overfitting the training data. A high variance means that the model performs exceptionally well on the training data as it has memorized it, but it fails to generalize to unseen testing data. Consequently, the model exhibits a significant drop in accuracy when evaluated on the testing data.