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 – 1: Using ridge regression, we can observe that when the magnitude of alpha climbs from zero, the error term decreases, and the train error shows a rising pattern as alpha rises. We opted to use a value of alpha equivalent to 2 for the ridge regression since the testing error is the lowest whenever the alpha value is 2.

Once we raise the level of alpha, the model tries to penalise more and attempt to make the majority of the coefficient values zero, therefore I opted to retain the alpha parameter at 0.01. Originally, the mean absolute error and alpha were both 0.4.

If we increase the value of alpha in my ridge regression by a factor of two, then we will consider the level of alpha to be equal to 10. This will cause the model to impose a greater penalties on the curve in an effort to render the model increasingly generalised, which will result in the model being made more straightforward and eliminate the need to consider how each individual data point fits into the overall picture.

When alpha is 10, we can observe that the test and the training both have higher errors when alpha is 10.

Like alpha for lasso, increasing r2 square likewise reduces the coefficient of the parameter to zero as we strive to penalise the model.

The following are the most critical variables for ridge regression once the modifications have been made:

x1	GrLivArea	0.128
x2	OverallQual	0.094
x3	OverallCond	0.061
x4	GarageArea	0.037

x5	TotalBsmtSF	0.036
x6	BsmtFinSF1	0.029
x7	Fireplaces	0.026
x8	LotArea	0.015
x9	LotFrontage	0.008
x10	BsmtFullBath	0.008

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 – 2: Predictive accuracy may be improved by regularising coefficients and rendering the model more readable as well as decreasing variance.

Ridge regression makes use of such a tuning parameter known as lambda, and the penalty for using this method is the square of the size of the coefficients. This is determined through cross validation. The penalty may be used to reduce the number of remaining sums or squares. In order to punish the coefficients with higher values, the cost is lambda times the sum of squared of the coefficients. The model's variance decreases, and its bias does not change when we raise lambda. In contrast to Lasso Regression, the Ridge Regression incorporates all variables into the final model.

Cross validation identifies a tuning parameter termed lambda as the cost in Lasso regression, which is used to adjust the size of the coefficients. Lambda grows, and as a result, the coefficient of Lasso reduces until it is precisely equal to 0. Lasso could also be used to pick variables. Smaller values of lambda provide basic linear regression; as lambda is larger, the model shrinks and ignores variables with 0 values.

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 – 3: The 5 most essential variables now are:

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

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 – 4: In order to increase robustness and generalizability, the model must be kept as basic as feasible. This will reduce accuracy. The Bias-Variance trade-off may also be used to explain it. There is greater bias in a basic model, but less variation and it is easier to generalise. To put it another way, the accuracy of a robust and generalizable model is unaffected by training or testing data; in other words, performance on training and testing data is almost same.

- Bias refers to the assumptions that the model makes in order to simplify things and make the target function simpler to estimate.
- A function's variance is the degree to which an approximation of a target function changes when new training data is introduced.
- The tension among bias-induced error and variation is what we mean by the term "trade-off."