UpGrad Advanced Regression Assignment

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 cross-validation, the optimal values of alpha were determined for ridge and lasso regression models:

1. Ridge Regression:

- Increased coefficient shrinkage towards zero but not exactly zero.
- Better management of multicollinearity.
- The importance rank may slightly change but remains relatively stable.

2. Lasso Regression:

- More coefficients driven to zero, resulting in a sparser model.
- Increased variable selection aggressiveness.
- The most important predictors may change significantly due to increased 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-2:

Choosing between ridge and lasso regression depends on the dataset and business needs:

a) Ridge Regression:

- Suitable when all predictors are potentially important.
- Handles multicollinearity well.
- Produces stable, less sparse models.

b) Lasso Regression:

- Performs variable selection by setting some coefficients to zero.
- Ideal when only a few predictors are significant.
- Produces sparser, more interpretable models.

Lasso regression is preferred in this case as it simplifies the model by selecting the most important predictors, improving interpretability and reducing the risk of overfitting.

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:

After the first five have been discarded, the next important sets of predictive variables may be determined again by re-running lasso regression. The following is an illustration of the idea:

1. Remove those five variables.

2. Take the modified dataset and redo a lasso model.

3. Get another set of significant predictors.

If for eg: v1, v2, v3, v4, v5 were found to be the top five predictors in the original lasso model, possible candidates for the most significant next predictor could be v6, v7, v8, v9, & v10. This next group will be based on non-zero coefficient values from this refitted lasso model.

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:

I will follow the below to ensure a good model quality:

1. Cross-Validation: Appraising the performance of training data on different subsets.
2. Regularization: Prevention of overfitting using ridge or lasso regression.
3. Feature Selection: Selection of features such as lasso regression or recursive feature elimination.
4. Ensemble Methods: Improving the model by combining multiple models.
5. Hyperparameter Tuning: Optimization of hyperparameters using grid or randomized search.
6. Data Augmentation and Noise Addition: Making changes in the training data.

Implications for Accuracy:

1. Increased Generalizability: Predictions that will be better for customers not seen, and will prevent over-fitting.
2. Potential Decrease in Training Accuracy: Robustness techniques usually decrease training accuracy but could improve test accuracy.
3. Improved Interpretability: Simpler models with fewer predictors are easier to interpret.