Assignment - 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:

The optimal value of alpha are :
Ridge Regression: 10.0
Lasso Regression: 0.001

If we double the value of alpha for both Ridge and Lasso Regression:
New alpha for Ridge: 20.0 (double of 10.0)
New alpha for Lasso: 0.002 (double of 0.001)

Changes in the Model:

For Ridge Regression:

- The regularization penalty would become stronger, leading to more shrinkage of coefficient values towards zero.
- The model would become more simpler as coefficients are further pushed towards zero.
- The overall effect of features might be more subdued, resulting in a smoother model.

For Lasso Regression:

- The regularization penalty would be intensified, causing more coefficients to be driven exactly to zero.
- The model would become even sparser as more features are removed from the model.
- The feature selection effect would become more pronounced, potentially resulting in a model with fewer predictors.

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:

The optimal values of lambda (alpha) determined for Ridge and Lasso Regression are:

Optimal alpha for Ridge: 10.0
Optimal alpha for Lasso: 0.001

The decision on which regularization technique to choose, Ridge or Lasso Regression, depends on the specific goals and the characteristics of the dataset.

Ridge Regression (alpha=10.0):

Ridge Regression helps to stabilize the coefficients and reduces model complexity. Ridge Regression might be preferred if we want a model that is stable and robust to small changes in the data.

Lasso Regression (alpha=0.001):
If the dataset has many features that are not relevant to the target variable and we want to perform feature selection, Lasso Regression is a good option. It tends to drive less important features to exactly zero, effectively removing them from the model. Lasso might be preferred if we want to identify a subset of features that have the strongest impact on the response variable.

Based on Performance:
If the primary goal is predictive accuracy, we can cross-validate both Ridge and Lasso models on our test data and select the one with the best performance in terms of metrics like R-squared, RMSE, or any other appropriate evaluation metric.

In this case, comparing the R2 scores and RMSE values, we can see the following:

R2 Scores: Both Lasso and Ridge Regression have similar R2 scores on the test set, with Lasso having a slightly higher R2 score. This suggests that both models explain a similar amount of variance in the test data.

RMSE Values: Both Lasso and Ridge Regression have similar RMSE values on the test set, with Ridge Regression having a slightly higher RMSE. This indicates that both models have a similar level of predictive accuracy on unseen data.

Based on the comparison, I have chosen **Lasso Regression** as my final model. It has slightly higher R2 score on the test set and lower RMSE, suggesting better predictive performance. Additionally, Lasso's ability for feature selection could be beneficial.

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 five most important predictor variables that are not available are:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. SaleCondition_Partial

To determine the five most important predictor variables in the new model after excluding the five most important predictor variables from the Lasso model, we need to do the following steps:

1. Train the Lasso Regression model on the original dataset with all features to identify the five most important predictor variables.
2. Determine the names or indices of these five important variables.
3. Create a new dataset by excluding these five important predictor variables.
4. Train a new Lasso Regression model on the modified dataset without these five variables.
5. Analyze the coefficients of the new Lasso model to identify the new five most important predictor variables.

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:

To make a model robust and generalizable, we can do the following:

- Cross-Validation: Use k-fold cross-validation to test performance on various data subsets.
- Split Data: Divide data into training and test sets for evaluating model accuracy.
- Feature Selection: Choose relevant features and preprocess data properly.
- Regularization: Apply Ridge or Lasso Regression to prevent overfitting.
- Hyperparameter Tuning: Optimize parameters for better generalization.
- Simplicity: Avoid overly complex models to reduce overfitting risk.
- Ensemble Methods: Combine models for more reliable predictions.
- Outlier Handling: Address outliers to prevent skewed predictions.
- Domain Knowledge: Incorporate expertise for informed decisions.

The implications of the same for the accuracy of the model would be :

- Robust models maintain consistent performance across datasets.
- Prioritizing generalization may trade some training accuracy.
- Striking a balance yields accurate predictions on new data.