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?

Ridge Regression: Increasing alpha in Ridge regression penalizes coefficients, but they will not become zero. Even if the coefficients shrink, all predictors tend to be retained in the model.

Lasso Regression: Lasso tends to perform feature selection by forcing some coefficients to become exactly zero. Doubling alpha in Lasso will further encourage sparsity, potentially reducing the number of predictors and emphasizing a subset of the most influential variables. The variables with non-zero coefficients after the change are considered the most important predictors.

Most important predictor variables after the change is implemented for Ridge Regression

10	MSZoning_RL	True	1	0.0851
5	GrLivArea	True	1	0.0763
1	OverallQual	True	1	0.0684
11	MSZoning_RM	True	1	0.0588
9	MSZoning_FV	True	1	0.0581
2	OverallCond	True	1	0.0453
4	TotalBsmtSF	True	1	0.0453

14	Foundation_PConc	True	1	0.0420
7	GarageCars	True	1	0.0359
3	BsmtFinSF1	True	1	0.0277

Most important predictor variables after the change is implemented for Lasso Regression

11	MSZoning_RL	True	1	0.106565
5	GrLivArea	True	1	0.100181
12	MSZoning_RM	True	1	0.076448
1	OverallQual	True	1	0.069656
9	MSZoning_FV	True	1	0.068896
4	TotalBsmtSF	True	1	0.046174
2	OverallCond	True	1	0.044992
14	Foundation_PConc	True	1	0.041865
7	GarageCars	True	1	0.036802
3	BsmtFinSF1	True	1	0.02857

Question 2

Assignment Part-II

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?

- The optimal lambda value in case of Ridge and Lasso is as below:

  - Ridge – 10

  - Lasso - 0.0004

- The Mean Squared error in case of Ridge and Lasso are:

  - Ridge - 0.013743

  - Lasso - 0.013556

- The Mean Squared Error of Lasso is slightly lower than that of Ridge

- Also, since Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge.

- Hence based on Lasso, the factors that generally affect the price are the Zoning classification, Living area

square feet, Overall quality and condition of the house, Foundation type of the house, Number of cars that can be accomodated in the garage, Total basement area in square feet and the Basement finished square feet area

Therefore, the variables predicted by Lasso in the above bar chart as significant variables for predicting the price of a house.

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?

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

1. GrLivArea

2. OverallQual

3. OverallCond

4. TotalBsmtSF

5. GarageArea

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?

Ensuring that a model is robust and generalizable involves several key steps:

Train-Test Split: Splitting the dataset into training and testing sets allows the model to learn patterns from the training data and validate its performance on unseen data (testing set). This helps assess how well the model generalizes to new, unseen instances.

Cross-Validation: Techniques like k-fold cross-validation further enhance model robustness by dividing the dataset into multiple subsets (folds) for training and testing. This method provides more reliable performance estimates by averaging the results across multiple iterations.

Feature Engineering and Selection: Careful feature engineering and selection can improve model generalization. Removing irrelevant or redundant features and creating meaningful features can enhance a model's ability to generalize well to new data.

Hyperparameter Tuning: Optimizing hyperparameters using techniques like grid search or random search ensures that the model's performance is not overly optimized for the training set but remains effective on unseen data.

Regularization: Applying regularization techniques (such as Ridge or Lasso regression) helps prevent overfitting by penalizing complex models, thereby enhancing their generalizability.

Handling Imbalanced Data: If the dataset is imbalanced, techniques like oversampling, undersampling, or using appropriate evaluation metrics (such as precision, recall, or F1-score) help ensure that the model's performance is not biased towards the majority class.

Implications for Model Accuracy:

Robust and generalizable models might sacrifice a small amount of accuracy on the training set. By ensuring a model's ability to generalize to new data, it might not perform as exceptionally well on the training set as an overfitted model would. However, it's more likely to perform better on new, unseen data.

A highly accurate model on the training set that lacks generalizability might suffer from overfitting. Such a model may perform poorly on new data due to its inability to capture underlying patterns and instead memorizes noise present in the training set. The ultimate goal is to strike a balance where the model performs well on both the training and testing sets, indicating that it has learned the underlying patterns without overfitting to the training data.