R.Kiruthiga

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?

Ans:

Optimal value of alpha for ridge and lasso regression

Ridge Regression:

```
print(ridge_model_cv.best_params_)

{'alpha': 2.0}
```

Lasso Regression:

```
print(lasso_model_cv.best_params_)

{'alpha': 10.0}
```

R2_score when the Alpha=2

```
R2_score training set 0.8771397084947814
R2_score test set 0.8676328501189295
```

When the alpha value is doubled i.e. Alpha = 4

```
R2_score training set 0.8771397084947814
R2_score test set 0.8676328501189295
```

Therefore no changes are observed when the alpha is doubled.

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?

Ans: I would prefer Ridge regression

Ridge regression introduces a regularization term proportional to the square of the coefficients' magnitude, implementing L2 regularization. The primary effect is a gradual shrinkage of coefficients towards zero, yet it typically avoids reducing any coefficient to precisely zero. This regularization technique is particularly advantageous when there is a belief that numerous features contribute to the model, and the aim is to dampen the influence of less crucial features without entirely discarding them. Ridge regression is particularly adept at mitigating the impact of multicollinearity, a scenario in which predictor variables exhibit high correlation.

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?

Ans:

Index(['OverallQual', 'TotRmsAbvGrd', 'GarageCars', 'GrLivArea',

'Neighborhood_NoRidge'],

dtype='object')

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?

Ans:

Machine learning models must be trained and evaluated using a variety of methodologies in order to ensure their robustness and generalizability. The term "robustness" describes the model's ability to function well on a variety of datasets, including ones it has never seen before. The capacity of the model to produce precise predictions on new, untested examples is shown by its generalizability. Achieving these objectives involves taking into account the following factors:

1. **Cross-Validation:**
    - Employ techniques such as k-fold cross-validation to gauge the model's performance across different data subsets. This minimizes the risk of overfitting to a specific subset and enhances overall robustness.
2. **Train-Test Split:**
    - Divide the dataset into distinct training and testing sets. Training the model on one subset and evaluating its performance on another simulates its effectiveness on new, previously unseen data.
3. **Validation Set:**
    - Introduce a validation set during training to assess the model's performance on data it has not encountered before. Adjusting hyperparameters based on this set helps prevent overfitting.
4. **Feature Engineering:**
    - Thoughtfully select and engineer features to ensure the model captures relevant patterns without fitting noise. Properly handling data through feature scaling, normalization, and addressing missing values contributes to a more robust model.
5. **Regularization:**
    - Implement regularization techniques, such as L1 or L2 regularization in linear models, to prevent overfitting and maintain an appropriate level of model complexity.

Implications for Accuracy:

Training Accuracy vs. Test Accuracy:

Discrepancies between high training accuracy and poor test accuracy may indicate overfitting. Ensuring robustness through cross-validation and proper evaluation helps align training and test performance.

Consistent Performance:

A model demonstrating consistent performance across different datasets signifies robustness and generalizability. Significant variations in accuracy between diverse data subsets may suggest issues with overfitting or a lack of generalization capability.