Assignment Part-II – Subjective Questions

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 ?**

Solution : Optimal value of lambda for Ridge Regression = 10
Optimal value of lambda for Lasso = 0.001

Below are the changes observed : (Code Base Attached in the Part 1 Python Notebook towards the end)

Changes in Ridge Regression metrics:
R2 score of train set decreased from 0.94 to 0.93
R2 score of test set remained same at 0.93

Changes in Lasso metrics:
R2 score of train set decreased from 0.92 to 0.91
R2 score of test set decreased from 0.93 to 0.91

The most important predictor variables after we double the alpha values are:-

1. GrLivArea
2. OverallQual_8
3. OverallQual_9
4. Functional_Typ
5. Neighborhood_Crawfor
6. Exterior1st_BrkFace
7. TotalBsmtSF
8. CentralAir_Y

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?**

Solution : Ridge and **Lasso** regression are commonly used machine learning techniques to impose constraints on model parameters.
Both methods aim to minimize the coefficient estimates towards zero, as doing so can significantly reduce variance, thereby addressing the issue of overfitting.

Ridge regression is a technique utilized in the analysis of multi-linear regression, particularly when dealing with multicollinearity. It is commonly referred to as L2 regularization. Ridge regression is applied when the predicted values exceed the observed values.

Ridge regression is employed when the objective is to prevent excessively large coefficients and prioritize the reduction of coefficient magnitudes.

Lasso, which stands for "Least Absolute Shrinkage and Selection Operator," is a technique that involves shrinking data points towards a central point, such as the mean. Lasso is also referred to as L1 regularization.
This technique is applied when the model is exhibiting overfitting issues or facing computational challenges.
If feature selection is a primary objective, particularly when dealing with a large number of variables, the Lasso technique is commonly employed.

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?**

Solution : New Important Predictor Variables now are (Code Base Attached in the Part 1 Python Notebook towards the end)

1. 2ndFlrSF
2. Functional_Typ
3. 1stFlrSF
4. MSSubClass_70
5. Neighborhood_Somerst

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?**

Solution : Robust and Generalizable mean as given below :
Robust Model – It refers to the behavior of a model which gets least affected when using new data/test data vs training data. Ideal expectation is the behavior should not deviate significantly.
Generalizable Model – It refers to the behavior of a model which exhibits effective adaptation to novel, previously unseen data, drawn from the same underlying distribution as the one used for model creation.
To **ensure both robustness and generalizability**, it is crucial to avoid overfitting. Overfitting, characterized by high variance, results in significant model prediction changes with even slight modifications in the data. While an overfitting model accurately captures all patterns in the training data, it fails to identify patterns in unseen test data.
Put simply, to achieve robustness and generalizability, the model should not be overly complex.

Considering **accuracy**, a highly complex model often yields superior accuracy. However, to enhance model robustness and generalizability, we need to reduce variance, which introduces some bias. Introducing bias reduces accuracy.
In general, striking a balance between model accuracy and complexity is essential. **Regularization techniques** such as **Ridge** Regression and **Lasso** can help achieve this balance.

END
