Assignment-based 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?

Answer 1

Optimal value for Ridge Regression: 0.05

Optimal value for Lasso Regression: 0.001

If Alpha value is doubled,

Ridge Regression:

- Doubling the alpha value in Ridge regression would increase the penalty for large coefficients.
- Consequently, the model will tend to shrink the coefficients more aggressively, pushing them closer to zero.
- This will lead to a simpler model with smaller coefficient values.

Top oefficients for Ridge:

PoolQC	1.366129
Street	0.137649
Utilities	0.093072
OverallQual	0.063523
CentralAir	0.055247
GarageQual	0.047058
OverallCond	0.046705
RoofMatl	0.046382
KitchenAbvGr	0.045240
BsmtFullBath	0.044795

Lasso Regression:

- Doubling the alpha value in Lasso regression would increase the penalty for large coefficients.
- This would lead to more coefficients being pushed exactly to zero.
- As a result, more features being effectively eliminated from the model.

Top coefficients for Lasso:

OverallQual	0.067282

RoofMatl	0.058674
OverallCond	0.043903
CentralAir	0.042452
GarageCars	0.038628
BsmtFullBath	0.037583
BsmtQual	0.029013
LandContour	0.025115
FullBath	0.024154.

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

Will choose the Lasso over Ridge as the R2 score and RMSE value of Lasso is better than Ridge on Train and Test data

```
Ridge R2_Train:  0.9069865489645849
Ridge R2_Test:  0.24754698327130198
------------------------------------
Lasso R2_Train:  0.895259069133675
Lasso R2_Test:  0.8480832350763765

------------------------------------

Ridge Regression:
Training MSE: 0.015195634725143308
Test MSE: 0.11325133626282377
------------------------------------
Lasso Regression:
Training MSE: 0.017111556538313515
Test MSE: 0.022864918135518525
```

- Training MSE: Ridge regression has a lower training MSE (0.0152) compared to Lasso regression (0.0171), indicating that Ridge regression fits the training data slightly better.
- Test MSE: Lasso regression has a lower test MSE (0.0229) compared to Ridge regression (0.1133), indicating that Lasso regression performs better on unseen data. This suggests that Lasso regression might be better at generalizing to new data.
- Overfitting: Ridge regression seems to be overfitting more than Lasso regression, as evidenced by the significant increase in test MSE compared to training MSE. Lasso's test MSE is closer to its training MSE, indicating better generalization.

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

5 most important predictor variables that will be excluded are

- OverallQual
- RoofMatl
- OverallCond
- CentralAir
- GarageCars

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

A robust and generalisable model will have consistent performance across different datasets, indicating its ability to capture underlying patterns in the data.

A well-behaved model should be unbiased, residuals should be randomly spread around 0, should be spreaded in a constant manner. The Bias and variance should be minimize to achieve optimal performance. High bias models make strong assumptions, leading to oversimplified representations of data (underfitting), while high variance models capture intricate patterns, including noise (overfitting). As model complexity increases to reduce bias, variance typically increases, and vice versa. The goal is to minimize both bias and variance to achieve optimal model performance on unseen data. Techniques such as regularization, cross-validation, and ensemble methods help manage this trade off by finding a suitable balance between bias and variance, ultimately improving the model's generalization capability.