Advanced Regression Assignment

Subjective question

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?

1. **What is the optimal value of alpha for ridge and lasso regression?**
    - The optimal value of alpha for ridge is 2 and lasso is 0.0004.
2. **What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?**
    - Doubling the value of alpha for both ridge and lasso increases the regularization strength, shrinking coefficients in Ridge, and more coefficients set to zero in Lasso.
    - But in practice, the value of alpha when doubling is still quite small so it does not have any significant changes
3. **What will be the most important predictor variables after the change is implemented?**

```
Ridge co-efficient
The most important predictor variables are as follows:
                      Important Alpha Co-Efficient
2ndFlrSF                                  0.136293
1stFlrSF                                  0.131856
TotalBsmtSF                               0.093411
Neighborhood_Crawfor                      0.078230
BsmtFinSF1                                0.055891
TotalArea                                 0.054394
GarageArea                                0.047950
Neighborhood_Veenker                      0.040872
MedNhbdArea                               0.040267
SaleCondition_Partial                     0.040142
Neighborhood_NridgHt                      0.038918
Condition1_Norm                           0.038569
MSZoning_RL                               0.038014
OpenPorchSF                               0.036035
MSZoning_FV                               0.034211
MasVnrArea                                0.033593
Exterior1st_BrkFace                       0.032437
Neighborhood_Somerst                      0.032158
Condition1_PosN                           0.029837
CentralAir_Y                              0.029562
```

	Ridge Doubled Alpha Co-Efficient
1stFlrSF	0.102410
2ndFlrSF	0.095409
TotalBsmtSF	0.077133
Neighborhood_Crawfor	0.070768
BsmtFinSF1	0.050777
TotalArea	0.047883
GarageArea	0.042884
MedNhbdArea	0.039198
Condition1_Norm	0.036869
Neighborhood_NridgHt	0.036484
OpenPorchSF	0.035897
MasVnrArea	0.031687
Neighborhood_Veenker	0.031065
Exterior1st_BrkFace	0.030571
Total_Bathrooms	0.030061
Neighborhood_Somerst	0.029537
SaleCondition_Partial	0.029470
YearRemodAdd	0.028763
CentralAir_Y	0.028396
MSZoning_RL	0.027866

```
Lasso co-efficient
The most important predictor variables are as follows:
                      Important Alpha Co-Efficient
1stFlrSF                                  0.188684
2ndFlrSF                                  0.185108
Neighborhood_Crawfor                      0.089685
TotalBsmtSF                               0.085700
BsmtFinSF1                                0.054460
MedNhbdArea                               0.044662
TotalArea                                 0.040130
GarageArea                                0.037743
Neighborhood_Somerst                      0.036702
Condition1_Norm                           0.033258
Neighborhood_NridgHt                      0.033065
Exterior1st_BrkFace                       0.030793
SaleCondition_Partial                     0.029245
OpenPorchSF                               0.028250
Exterior1st_CemntBd                       0.027337
GarageCars                                0.024443
Functional_Typ                            0.022864
Total_Bathrooms                           0.022194
OverallCond                               0.019632
YearRemodAdd                              0.018820
```

	Lasso Doubled Alpha Co-Efficient
1stFlrSF	0.153345
2ndFlrSF	0.145544
Neighborhood_Crawfor	0.081604
TotalBsmtSF	0.072960
BsmtFinSF1	0.049909
Exterior1st_CemntBd	0.033792
GarageCars	0.029880
TotalArea	0.029820
GarageArea	0.029118
Condition1_Norm	0.028848
MedNhbdArea	0.028121
OpenPorchSF	0.028032
Total_Bathrooms	0.026267
Neighborhood_NridgHt	0.024717
Total_Home_Quality	0.022518
Exterior1st_BrkFace	0.022329
YearRemodAdd	0.022309
SaleCondition_Partial	0.020681
Neighborhood_Somerst	0.019731
CentralAir_Y	0.018814

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which on will you choose to apply and why?

With Ridge regression:

- The optimum alpha is 2.0
- The R2 Score of the model on the test dataset for optimum alpha is 0.8752062671978895
- The MSE of the model on the test dataset for optimum alpha is 0.003967385658240671

With Lasso regression:

- The optimum alpha is 0.0004
- The R2 Score of the model on the test dataset for optimum alpha is 0.8754907805311707
- The MSE of the model on the test dataset for optimum alpha is 0.003958340539606166

Both R2 score and MSE are almost the same.
Based on the business goal “The company wants to know which variables are significant in predicting the price of a house”. I chose to use Lasso to apply because It helps with feature reduction.

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?

- The five most important predictor variables: '1stFlrSF','2ndFlrSF','Neighborhood_Crawfor','TotalBsmtSF','BsmtFinSF1' are illustrated in the figure in question 1
- After remove them and train again with alpha = 0.0004, the new five most important predictor variables are:

	Lasso Co-Efficient
MedNhbdArea	0.119675
Exterior1st_CemntBd	0.075143
TotalArea	0.073713
GarageArea	0.059214
Total_Bathrooms	0.045767

Question 4

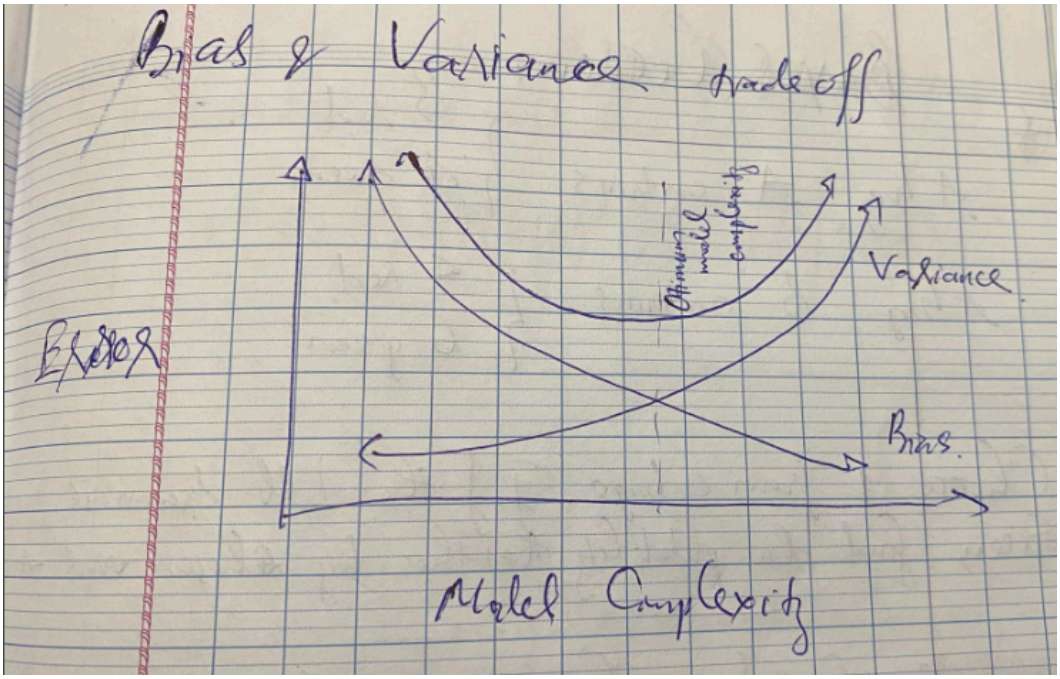
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

1. **How can you make sure that a model is robust and generalisable?**

Based on my assignment, I make sure that a model is robust and generalisable by:

a. Data Cleaning: Address missing values, outliers, and anomalies in your dataset.
b. Feature Scaling: Standardize or normalize numerical features to ensure that all features contribute equally to the model.
c. Identify and select relevant features for your model. Remove irrelevant or redundant features that do not contribute significantly to the predictive power.
d. Check for multicollinearity among independent variables. High multicollinearity can affect the stability of coefficient estimates. Techniques like variance inflation factor (VIF) can be used to detect and address multicollinearity.

e. Use techniques like k-fold cross-validation to assess how well the model generalizes to new data. This helps in detecting overfitting or underfitting issues.
f. Apply regularization techniques such as Lasso or Ridge regularization to prevent overfitting and improve model generalization.
g. Choose appropriate evaluation metrics (e.g., mean squared error, R-squared) to assess the model's performance.
h. Split your dataset into training and testing sets to evaluate the model's performance on unseen data.
i. Making a model simple lead to Bias-Variance trade off.



2. **What are the implications of the same for the accuracy of the model and why?**

   The implications of these strategies for accuracy lie in their collective ability to enhance the model's reliability, reduce overfitting, handle outliers, and improve generalization to new, unseen data. By addressing these aspects, I increase the likelihood that the model will make accurate predictions across a broader range of scenarios.