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:

In Ridge and Lasso Regression, the alpha parameter controls the strength of regularization. Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function. The optimal value of alpha is typically chosen through cross-validation, where different values are tested to find the one that minimizes the prediction error on unseen data.

Optimal Alpha Values:

- For Lasso Regression, the optimal value of alpha is **0.001**.
- For Ridge Regression, the optimal value of alpha is **100.**

Changes in Model with Double Alpha:

For Ridge Regression (with alpha doubled to 200):

```
The R2 Score of the model on the test dataset for doubled alpha is 0.920231860337209
The MSE of the model on the test dataset for doubled alpha is 0.010285819764144194
The most important predictor variables are as follows:
```

	Ridge Doubled Alpha Co-Efficient
OverallQual	0.040513
GrLivArea	0.039590
TotalBsmtSF	0.030539
1stFlrSF	0.027473
BsmtFinSF1	0.023959

For Lasso Regression (with alpha doubled to 0.002):

```
The R2 Score of the model on the test dataset for doubled alpha is 0.9193196525138927
The MSE of the model on the test dataset for doubled alpha is 0.010403445739850024
The most important predictor variables are as follows:
```

	Lasso Doubled Alpha Co-Efficient
GrLivArea	0.122116
YearBuilt	0.079026
OverallQual	0.056099
OverallCond	0.038467
TotalBsmtSF	0.036380

Most Important Predictor Variables after Changes in Lasso Model (Excluding Top 5):

After excluding the five most important predictor variables (GrLivArea, YearBuilt, OverallQual, OverallCond, TotalBsmtSF) from the incoming dataset and building a new Lasso model:

```
The R2 Score of the model on the test dataset is 0.910395160483569
The MSE of the model on the test dataset is 0.011554227454184898
The most important predictor variables are as follows:
```

	Lasso Co-Efficient
2ndFlrSF	0.115959
1stFlrSF	0.090752
BsmtFinSF1	0.082599
BsmtUnfSF	0.048117
YearRemodAdd	0.043341

The R2 Score and MSE values provide insights into the performance of the models. A higher R2 Score indicates better model performance, while a lower MSE is desirable for better accuracy. The predictor variables are listed with their corresponding coefficients, indicating their influence on the target variable.

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:

I would prefer Lasso Regression.

Lasso has a tendency to shrink the coefficients of less important features toward zero, effectively performing feature selection by assigning zero coefficients to some variables. This helps in identifying and emphasizing the most relevant predictors.

If multicollinearity is a concern: Ridge Regression might be a better choice. Ridge regression handles multicollinearity well by shrinking the coefficients of correlated variables without necessarily eliminating any. It tends to distribute the weight across all correlated variables.

In summary: we have to use Lasso Regression if you want a sparse model with feature selection. Or Use Ridge Regression if you want to mitigate multicollinearity and maintain all features in the model. The final decision depends on the specific characteristics of the data and the goals of the analysis, based on the analysis I would prefer Lasso.

Question 3

After building the model, you realized 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?In a Lasso regression model, the importance of predictor variables is determined by the magnitude of their corresponding coefficients. When you exclude the five most important predictor variables, the next set of most important variables will be those with the highest absolute coefficients in the current model.

Answer:

```
#Removing the 5 most important predictor variables from the incoming dataset
X_test_rfe = X_test.drop(['GrLivArea','YearBuilt','OverallQual','OverallCond','TotalBsmtSF'],axis=1)
X_train_rfe = X_train.drop(['GrLivArea','YearBuilt','OverallQual','OverallCond','TotalBsmtSF'],axis=1)

# Building Lasso Model with the new dataset
lasso3 = Lasso(alpha=0.001,random_state=100)
lasso3.fit(X_train_rfe,y_train)
lasso3_coef = lasso3.coef_
y_test_pred = lasso3.predict(X_test_rfe)
print('The R2 Score of the model on the test dataset is',r2_score(y_test, y_test_pred))
print('The MSE of the model on the test dataset is', mean_squared_error(y_test, y_test_pred))
lasso3_coeff = pd.DataFrame(np.atleast_2d(lasso3_coef),columns=X_train_rfe.columns)
lasso3_coeff = lasso3_coeff.T
lasso3_coeff.rename(columns={0: 'Lasso Co-Efficient'},inplace=True)
lasso3_coeff.sort_values(by=['Lasso Co-Efficient'], ascending=False,inplace=True)
print('The most important predictor variables are as follows:')
lasso3_coeff.head(5)
```

```
The R2 Score of the model on the test dataset is 0.910395160483569
The MSE of the model on the test dataset is 0.011554227454184898
The most important predictor variables are as follows:
```

	Lasso Co-Efficient
2ndFlrSF	0.115959
1stFlrSF	0.090752
BsmtFinSF1	0.082599
BsmtUnfSF	0.048117
YearRemodAdd	0.043341

1. **2ndFlrSF (Second Floor Square Feet):** This variable has a significant positive impact on the predicted outcome. Properties with larger second-floor areas are associated with higher predictions.

2. **1stFlrSF (First Floor Square Feet):** The size of the first-floor living area has a substantial positive influence on the predicted outcome. Larger first-floor areas are correlated with higher predicted values.

3. **BsmtFinSF1 (Type 1 Finished Square Feet of Basement):** The finished square footage of the basement (BsmtFinSF1) is an essential predictor, positively contributing to the predicted outcome.

4. **BsmtUnfSF (Unfinished Square Feet of Basement):** The size of the unfinished basement area (BsmtUnfSF) positively influences the predicted outcome. Properties with larger unfinished basements tend to have higher predictions.

5. **YearRemodAdd (Remodeling Date):** The year of remodeling has a positive effect on the predicted outcome. Properties that underwent remodeling in specific years are associated with higher predictions.

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:

To ensure a model is robust and generalizable:

1. **Cross-Validation:**By techniques like k-fold cross-validation to assess model performance on multiple subsets of the data. This helps us to evaluate how well the model generalizes to different datasets.
2. **Feature Engineering:** Choosing relevant features and avoiding overfitting by removing unnecessary variables. For regression models we have RFE recursive feature dropping based and VIF Feature engineering helps create a more generalized model.
3. **Regularization Techniques:**Apply regularization methods like L1 or L2 regularization to prevent overfitting. These techniques penalize complex models, encouraging simplicity and generalization.
4. **Hyperparameter Tuning:**Fine-tune model hyperparameters to optimize performance. Use techniques like grid search to find the best parameter values for robustness.

Implications for Model Accuracy:

1. **Trade-off with Complexity:**A more robust and generalizable model may sacrifice some accuracy on the training data. It aims for a balance between accuracy and the ability to generalize to new, unseen data.
2. **Avoiding Overfitting:**Robust models are less prone to overfitting, ensuring they don't memorize the training data but learn patterns that can be applied to new data.
3. **Reduced Sensitivity to Noise:**Robust models are less affected by noise or random fluctuations in the training data, leading to more consistent and reliable predictions.