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

a) Optimal Value of Alpha:

- Optimal value of lambda for Ridge Regression = 4
- Optimal value of lambda for Lasso = **0.001**

b) Changes in the model if you choose double the value of alpha:
(Please refer the jupyter (.ipynb) file for the code. Results are mentioned below)

Ridge Regression: Original Model (alpha=4), Doubled Alpha Model(alpha=8)
Lasso Regression: Original Model (alpha=0.001), Doubled Alpha Model(alpha=0.002)

i. Observation on Performance metrics:

Metric	Ridge Regression	Ridge Regression_doubled	Lasso Regression	Lasso Regression_doubled
R2 Score (Train)	0.917	0.913	0.914	0.907
R2 Score (Test)	0.903	0.905	0.903	0.906
RSS (Train)	84.358	89.127	87.654	95.371
RSS (Test)	43.776	42.771	43.838	42.496
MSE (Train)	0.083	0.087	0.086	0.093
MSE (Test)	0.100	0.098	0.100	0.097
RMSE (Train)	0.287	0.295	0.293	0.306
RMSE (Test)	0.316	0.312	0.316	0.311

Ridge Regression:
In general, the primary goal is to have a model that generalizes well to new data. Therefore, the model with alpha = 8 seems to be better in terms of test data performance, as it shows improved R-squared, reduced RSS, and RMSE on the test data, while the changes in training data metrics are relatively small.

- Doubling the alpha from 4 to 8 has slightly reduced the R-squared on the training data, indicating a less accurate fit to the training data.
- However, the R-squared on the test data remains the same, indicating that the model's ability to generalize to new data is not significantly affected.
- The RSS on the training data has increased, indicating larger errors, while the RSS on the test data has decreased, indicating smaller errors.
- The MSE on the training data has increased, but the MSE on the test data remains the same.
- The RMSE on the training data has increased slightly, but the RMSE on the test data remains the same.

The model with alpha = 4 has a better fit to the training data (higher R-squared) compared to the model with alpha = 8. However, the difference is slight.The model with alpha = 8 performs slightly better on the test data, as indicated by the increased R-squared and reduced RSS and RMSE. This suggests that the model with higher alpha generalizes better to new, unseen data.

Lasso Regression:

Doubling the alpha from 0.001 to 0.002 has a minor impact on the model's performance. It slightly reduces overfitting, as indicated by the improvement in R-Squared on the test data and the decrease in RSS on the test data. However, it also results in a slight increase in RSS on the training data.

ii. Observation on the magnitude of coefficients:

For majority of cases more shrinkage of coefficient values are observed on doubling the alpha for both ridge and lasso regression

	Ridge	Ridge_Doubled	Lasso	Lasso_Doubled
TotalBsmtSF	0.166	0.170	0.170	0.171
LowQualFinSF	0.000	0.000	0.000	0.000
GrLivArea	0.336	0.344	0.336	0.354
GarageCars	0.140	0.147	0.137	0.144
3SsnPorch	0.000	0.000	0.000	0.000
MiscVal	0.000	0.000	0.000	0.000
AgeofProperty	-0.167	-0.169	-0.163	-0.171
MSSubClass_40	0.064	0.040	0.000	0.000
MSSubClass_60	0.150	0.127	0.112	0.062
MSSubClass_70	0.119	0.107	0.047	0.000
MSSubClass_75	0.191	0.137	0.145	0.030
MSSubClass_90	-0.211	-0.199	-0.432	-0.409
MSSubClass_120	-0.117	-0.101	-0.000	-0.000
MSSubClass_160	-0.103	-0.108	-0.000	-0.039
MSSubClass_180	-0.107	-0.081	-0.000	-0.000
MSSubClass_190	-0.029	-0.038	-0.000	-0.000
MSZoning_FV	0.229	0.194	0.222	0.166
MSZoning_RH	0.030	0.004	0.000	-0.000
LotShape_IR2	0.134	0.123	0.117	0.098

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

Ridge	Lasso
OverallQual_9_Excellent	OverallQual_9_Excellent
OverallQual_8_VeryGood	OverallQual_8_VeryGood
GrLivArea	OverallCond_9_Excellent
OverallCond_9_Excellent	MSSubClass_90
Neighborhood_Crawfor	Neighborhood_Crawfor
SaleCondition_Alloca	GrLivArea
OverallCond_3_Fair	BldgType_Twnhs
Neighborhood_NridgHt	Exterior2nd_BrkFace
Neighborhood_NoRidge	OverallCond_7_Good
KitchenQual_TA	SaleCondition_Alloca

Coefficient details are given below

Ridge:

	Features	Coefficient	Abs_Coefficient_Ridge(Desc_Sort)
0	OverallQual_9_Excellent	0.511	0.511
1	OverallQual_8_VeryGood	0.417	0.417
2	GrLivArea	0.344	0.344
3	OverallCond_9_Excellent	0.324	0.324
4	Neighborhood_Crawfor	0.316	0.316
5	SaleCondition_Alloca	0.262	0.262
6	OverallCond_3_Fair	-0.246	0.246
7	Neighborhood_NridgHt	0.244	0.244
8	Neighborhood_NoRidge	0.240	0.240
9	KitchenQual_TA	-0.238	0.238

Lasso:

	Features	Coefficient	Abs_Coefficient_Lasso(Desc_Sort)
0	OverallQual_9_Excellent	0.724	0.724
1	OverallQual_8_VeryGood	0.501	0.501
2	OverallCond_9_Excellent	0.427	0.427
3	MSSubClass_90	-0.409	0.409
4	Neighborhood_Crawfor	0.355	0.355
5	GrLivArea	0.354	0.354
6	BldgType_Twnhs	-0.242	0.242
7	Exterior2nd_BrkFace	0.240	0.240
8	OverallCond_7_Good	0.218	0.218
9	SaleCondition_Alloca	0.216	0.216

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

Performance metrics is given below

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.917	0.914
R2 Score (Test)	0.903	0.903
RSS (Train)	84.358	87.654
RSS (Test)	43.776	43.838
MSE (Train)	0.083	0.086
MSE (Test)	0.100	0.100
RMSE (Train)	0.287	0.293
RMSE (Test)	0.316	0.316

Observation:

There is no substantial difference in the model performance between Ridge and Lasso regression techniques. However, my preference leans toward employing Lasso in this specific use case due to its inherent feature selection capabilities, which promote the selection of relevant features while maintaining coefficients at a manageable scale, thereby contributing to the model's overall performance.

Ridge Regression, on the other hand, is less suitable for feature selection and becomes less practical when dealing with a high-dimensional dataset with over 200 features. In such cases, it is advisable to explore alternative feature selection approaches in conjunction with Ridge regression to address the issue of high coefficient magnitudes.

1. If our main goal is to achieve model simplicity and reduce the features by setting some coefficients to zero, Lasso Regression is a more appropriate choice.

Ridge Regression is often employed when you believe that all features are relevant to some extent and want to avoid feature selection or when you have no prior knowledge about the importance of specific features.

2. If multicollinearity is a concern and if we aim to reduce the magnitude of coefficients while retaining all features, Ridge Regression is a good choice. This doesn't force coefficients to zero and can handle correlated features more gracefully by stabilizing the model and improving its generalization.
3. The choice between Ridge and Lasso is data-driven and based on which method results in better predictive performance. Hence the final decision should be validated through cross-validation, as we need to strike a balance between different objectives -Trade-off Between Overfitting and Coefficient Shrinkage, to ensure it's the best choice for our specific dataset

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

(Please refer the jupyter (.ipynb) file for the code. Results are mentioned below)

Top five features in Lasso Model (before removing) were as follows:

```
0    OverallQual_9_Excellent
1    OverallCond_9_Excellent
2       SaleCondition_Alloca
3     OverallQual_8_VeryGood
4              MSSubClass_90
```

Top five predictor variables in the new model: (After removing the above top 5 predictors):

```
0         BldgType_Duplex
1      OverallCond_3_Fair
2    Neighborhood_StoneBr
3    Neighborhood_Crawfor
4          KitchenQual_TA
```

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

Steps to ensure that a model is robust and generalizable is given below

1. Use adequate amount of high-quality and diverse data
2. Feature engineering- Select and engineer relevant features, eliminating irrelevant or redundant ones. Carefully crafted features can enhance model performance.
3. Implement cross-validation to assess the model's performance on various data subsets
4. Apply regularization techniques like Ridge or Lasso to prevent overfitting and improve generalization.
5. Fine-tune hyperparameters to optimize model performance, but do so using a separate validation set to avoid overfitting.
6. Assess the model's performance on a separate, unseen test dataset to gauge how well it generalizes to new data.

By following these practices, a model becomes more robust, meaning it is less sensitive to small changes in the training data and is more likely to perform consistently well across different datasets. A robust model is more likely to generalize effectively to unseen data, improving its predictive accuracy in real-world scenarios. Aim is to strike a balance between accuracy and the ability to handle new, unseen data effectively. While robustness and generalizability are essential, there may be a trade-off with model accuracy during training. Over-regularization or feature reduction can lead to reduced training accuracy, but this can be an acceptable trade-off if it results in a more generalizable model with better real-world performance**.**

Accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.