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?

Answer1

The optimal value of alpha for ridge regression: 10.

The optimal value of alpha for lasso regression: 0.001.

When changing the alpha for ridge regression to 20:

Alpha of Ridge	R2 Score for training	R2 Score for testing
10	0.928	0.874
20	0.918	0.872

The important predictor variables change is shown below:

Alpha of Ridge = 10

	Feature	Coeff
121	OverallCond_2	0.323981
68	Neighborhood_Edwards	0.234719
78	Neighborhood_OldTown	0.211106
120	OverallQual_9	0.192754
129	RoofStyle_Gable	0.183495

Alpha of Ridge = 20

	Feature	Coeff
121	OverallCond_2	0.251700
68	Neighborhood_Edwards	0.187100
78	Neighborhood_OldTown	0.176119
120	OverallQual_9	0.170500
12	BsmtFullBath	0.169240

When changing the alpha for lasso regression is 0.002:

Alpha of Lasso	R2 Score for training	R2 Score for testing
0.001	0.947	0.736
0.002	0.928	0.815

The important predictor variables change is shown below:

Alpha of Lasso = 0.001	Alpha of Lasso = 0.002

	Feature	Coeff
121	OverallCond_2	0.497116
12	BsmtFullBath	0.334736
129	RoofStyle_Gable	0.307779
68	Neighborhood_Edwards	0.292590
120	OverallQual_9	0.259716

	Feature	Coeff
121	OverallCond_2	0.487973
12	BsmtFullBath	0.329861
68	Neighborhood_Edwards	0.272055
120	OverallQual_9	0.260247
83	Neighborhood_StoneBr	0.214038

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?

Answer2

I will choose Ridge Regression here.

Both Ridge and Lasso regression have their own pros and cons. Ridge regression prevents overfitting well but does not force any coefficients to exactly zero. Lasso regression can perform feature selection by driving some coefficients to exactly zero which makes the model more interpretable, especially when dealing with a large number of features.

Here, in this case, we found that Lasso with the best lambda seems to face the problem of overfitting in the training dataset and we don't need to integrate the model, so we choose Ridge regression for the final decision.

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?

Answer3

After dropping the 5 top important predictor variables, the 5 new most important predictor variables are:

	Feaure	Coef
12	BsmtHalfBath	0.329140
125	RoofStyle_Gambrel	0.314747
67	Neighborhood_Gilbert	0.298431
81	Neighborhood_StoneBr	0.204397
76	Neighborhood_OldTown	0.178002

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?

Answer4

There are several methods to make a model robust and generalized:

1) **Train-Test Split**: Train the model on the training data and evaluate its performance on the testing data to assess generalization capability.
2) **Cross-Validation**: Utilize k-fold cross-validation to validate the model's performance on multiple subsets of the data.
3) **Feature Selection**: Select relevant features and remove irrelevant ones based on domain knowledge to prevent overfitting.
4) **Regularization**: Ridge adds an L2 penalty, and Lasso adds an L1 penalty to the loss function to prevent overfitting.

5) **Hyperparameter Tuning**: Optimize hyperparameters using techniques like grid search or random search.
6) **Outlier Handling**: Address outliers properly as they can impact model performance.
7) **Bias-Variance Tradeoff**: Aim to find the right balance between bias and variance to achieve a model that generalizes well.

Implications for Model Accuracy:

A robust and generalizable model shows consistent performance on different subsets of the data during cross-validation.

For Ridge and Lasso, regularization helps prevent overfitting, leading to better accuracy on new, unseen data.

If a model is not generalizable, it may have a high training accuracy but perform poorly on new data, resulting in overfitting and reduced predictive capabilities.

By implementing proper validation techniques, feature selection, regularization, and hyperparameter tuning, we can build models that perform well on new data and provide accurate predictions in practical applications.