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:

The optimal value depends on the specific data and problem being addressed to. Alpha is a hyperparameter that controls the strength of the regularization in the models, and it is what help in preventing overfitting by penalizing large coefficients.

The optimal values of alpha can be found using cross validation, where different values of alpha are tried and the one that gives best performance is selected.

If we double the value of alpha for both lasso and ridge , the models coefficients will be further penalized, which will lead to a large reduction in magnitude of the coefficients. This will result is reduction of the model complexity and might underfit the data.

The most important predictor variables after the change is implemented will depend on the data and the values of other hyperparameters in the model. The most important variable will have largest absolute values.

Question 2:

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

Answer: In the use case, we have already completed the feature selection and intended to keep all of the values. There are predictors that are highly correlated and since ridge regression shrinks the coefficients of correlated predictors towards each other, leading to more stable and reliable estimates. We use Ridge regression.

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 model excluding the five most important predictor variables. Which are the five most important variables now.

Answer:

The top five predictors are as follows:

	Ridge (alpha=20)	Lasso (alpha=500)
MSSubClass	3530.847673	1574.270039
LotFrontage	1951.951484	981.467069
LotArea	4217.936678	3598.031393
OverallQual	6980.906014	6691.723892
OverallCond	3787.864556	3767.075037

If these are removed, then the model might lose its performance , with reduced R^2.

Question 4:

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

Answer:

To ensure that a model is robust and generalizable, several steps can be taken, including:

1. Using a representative and diverse dataset: The dataset used to train and evaluate the model should be representative of the population it is meant to generalize to, and it should contain a diverse range of samples. This helps ensure that the model is not biased towards any particular subset of the data and can generalize well to new and unseen data.

2. Using cross-validation: Cross-validation is a technique that can help assess the model's performance on new and unseen data by testing it on different subsets of the data during training. This helps ensure that the model is not overfitting to the training data and can generalize well to new data.

3. Regularization: Regularization techniques such as L1 or L2 regularization can help prevent overfitting and improve the model's ability to generalize to new data.

4. Tuning hyperparameters: Hyperparameters such as learning rate, regularization strength, and number of layers in a neural network can significantly impact the model's generalizability. Tuning these hyperparameters using techniques such as grid search or random search can help identify the optimal values that lead to better generalization.

5. Testing the model on a holdout dataset: A holdout dataset that is completely separate from the training and validation data can be used to test the model's generalization ability. This helps ensure that the model can perform well on new and unseen data.

The implications of ensuring a model is robust and generalizable are that it can perform well on new and unseen data, rather than just on the training data. This means that the model is more likely to be useful in practice and can make accurate predictions on new data that it has not seen before. However, it's important to note that ensuring a model is robust and generalizable can sometimes come at the cost of accuracy on the training data. This is because techniques like regularization can reduce the model's ability to fit the training data precisely, in favor of better generalization to new data. Therefore, a trade-off between accuracy on the training data and generalization to new data may need to be made.