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:

The optimal value of alpha for ridge and lasso regression depends on the specific dataset and the desired trade-off between model complexity and the amount of regularization. Alpha is an hyperparameter that controls the amount of regularization applied in these regression techniques.

In ridge regression, the optimal value of alpha is typically determined using cross-validation techniques. A higher value of alpha increases the amount of regularization, resulting in a more constrained model with smaller coefficient values.

In lasso regression, the optimal value of alpha also depends on the dataset. Lasso regression has the property of performing variable selection by setting some coefficients to exactly zero. Higher values of alpha increase the amount of regularization and tend to set more coefficients to zero, leading to sparser models.

If you choose to double the value of alpha for both ridge and lasso regression, the models will become more regularized and more constrained. The coefficients will be further shrunk towards zero, potentially reducing the impact of some predictors.

After implementing the change, the most important predictor variables will depend on the specific dataset and the nature of the relationships within it. However, generally speaking, the most important predictor variables after increasing alpha would be those that have the highest absolute coefficient values in the updated models. These variables are the ones that have a relatively stronger influence on the model's predictions, even after the increased regularization. It is important to note that the impact on predictor variables can vary depending on the dataset, and it is recommended to perform a thorough analysis and evaluation to determine the most important variables in a given context.

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:

Though Ridge Regression performed better in terms of R2 values of Train and Test, Lasso is preferable because it brings and assigns a zero value to irrelevant features, allowing us to choose the predictive variables.
It is usually best to choose a simple but sturdy model.

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:

If the five most important predictor variables in the lasso model are not provided in the incoming data, it means that those variables cannot be included in the subsequent model. In this case, we would need to identify the next set of important predictor variables from the remaining available ones.

To determine the five most important predictor variables in the new model, we can follow these steps:

1. Fit the Lasso regression model using the available predictor variables.
2. Identify the coefficients of the predictor variables in the model.
3. Rank the predictor variables based on the absolute magnitude of their coefficients.
4. Select the five predictor variables with the highest absolute coefficient values as the most important in the new model.

It's important to note that the importance of predictor variables can vary depending on the specific dataset and the modeling techniques used. Therefore it is recommended to perform this process on your actual data to obtain the updated set of important predictor variables for our specific situation.

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 that a model is robust and generalizable, there are several steps we can take:

1. **Use Sufficient and Diverse Data**: Gather a sufficiently large and representative dataset that covers a wide range of scenarios and variations present in the real world. A larger and more diverse dataset can help the model capture a wide range of patterns and make it more robust.
2. **Split Data for Training and Testing**: Divide the dataset into separate training and testing sets. The training set is used to build and optimize the model, while the testing set is used to evaluate its performance on unseen data. This helps assess the model's ability to generalize to new data.
3. **Cross-Validation**: Employ techniques such as k-fold cross-validation to further evaluate the model's performance. This involves splitting the data into multiple subsets, training the model on different combinations of subsets, and assessing its performance on the remaining subset. Cross-validation provides a more comprehensive assessment of the model's generalization capabilities.
4. **Regularization**: Apply regularization techniques like ridge regression or lasso regression to prevent overfitting, which occurs when the model memorizes the training data but fails to

generalize to new data. Regularization helps control model complexity and improves generalization.

5. **Feature Engineering and Selection**: Carefully engineer features or select relevant features that are likely to have a meaningful impact on the target variable. Removing irrelevant or redundant features can reduce noise in the data and improve the model's generalization.
6. **Avoid Over-Optimization**: Avoid excessive tuning of hyperparameters based on the testing set performance alone. Using a separate validation set or nested cross-validation can help prevent over-optimization and provide a more accurate assessment of the model's generalization performance.
7. **Evaluate on Unseen Data**: Once the model is trained and optimized, evaluate its performance on completely unseen data, such as a holdout dataset or real-world data. This provides an important measure of the model's ability to generalize beyond the training and testing sets.

The implications of ensuring model robustness and generalizability for the accuracy of the model are that it might not achieve the highest accuracy on the training set alone. By prioritizing generalization, the model may sacrifice some accuracy on the training set in order to perform better on unseen data. This trade-off aims to prevent overfitting and increase the model's applicability to new and unseen instances. Ultimately, the goal is to have a model that performs well not just on the training set but also on real-world data, making it more reliable and trustworthy.