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 optimum value of alpha in Ride was 7 while in Lasso was 0.001. The choice of alpha value in ridge and lasso regression determines the objective of the model. A value of alpha equal to zero results in the same objective as linear regression, while an alpha value approaching infinity shrinks the coefficients towards zero. This is due to the infinite weightage placed on the square of the coefficients. Consequently, any value less than zero makes the objective function infinite.

The magnitude of alpha determines the relative importance given to different parts of the objective function. For simple linear regression, the coefficients are either zero or one. The optimal value of alpha is chosen based on the specific purpose of the analysis. Ridge and lasso regularization are designed to penalize model complexity. Higher values of alpha lead to less complex models, reducing the error due to variance and overfitting. Conversely, excessively high alpha values increase the error due to bias and underfitting.

Selecting the optimal alpha is crucial to minimize errors in both directions. It involves finding a balance between model complexity and generalization performance, ultimately achieving the best trade-off between bias and variance.

When the alpha value is doubled in ridge and lasso regression, several changes are observed in the model. The coefficients of each feature variable are decreased, leading to a decrease in both Y-train and Y-test predictions. Additionally, there is an increase in the residual sum of squares (RSS), AIC, and BIC values, indicating a decrease in model performance.

After doubling the alpha value, the most critical predictor variables may change. Variable selection is reduced by approximately 68% compared to the optimal alpha conditions. For example, when alpha is at its optimum, 79 features are selected, but when the alpha is doubled, the number of selected features reduces to 54. The priority and importance of predictor variables can also change with the increased alpha value.

It's important to note that while statistical measures can provide insights into the relative importance of predictor variables, determining their practical significance requires domain knowledge and context-specific understanding. Practical relevance and importance of variables cannot solely be determined through statistical measures. Therefore, it is crucial to incorporate domain expertise and consider the specific context when interpreting the results and selecting predictor variables.

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:

In predicting the sale price of houses using Lasso regression with an alpha value of 0.001 and Ridge regression with an alpha value of 50, various key predictor variables and their subcategories have been identified. These variables include Neighborhood_NridgHt, Neighborhood_Somerst, Neighborhood_StoneBr, Neighborhood_NoRidge, and many others. In total, 77 categories have been determined as significant predictors of house sales price. These variables collectively account for an observed R-squared value of approximately 87%, indicating their ability to explain the variation in house prices.

Additionally, the model's goodness of fit and simplicity are evaluated using the AIC and BIC criteria. The AIC value is reported as 2471, reflecting the balance between model fit and complexity. Similarly, the BIC value is given as 2661. These criteria help assess the trade-off between overfitting and underfitting in the model. Lasso Regression would be best in this senario.

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:

To enhance the modeling approach, the year value data columns were converted into categorical values. This conversion allows for a more effective representation of the relationship between the year values and the target variable. Additionally, the model was rebuilt to incorporate these categorical variables.

After the conversion, the five key variables that emerged as significant predictors of the house sales price were Neighborhood with subcategories including Somerset, NridgHt, and Crawford, BsmtFullBath, and Condition2 with the subcategory PosN.

By converting the year value data into categorical variables, the model can capture any non-linear relationships or patterns specific to different time periods. This transformation provides a more nuanced understanding of the impact of the neighborhood, basement full bathrooms, and the condition of the property on the house sales price.

Furthermore, it is important to note that binary categorical data was also converted in the process. This conversion allows the model to account for the influence of specific binary variables on the target variable. The identification of significant variables, such as Somerset, NridgHt, Crawford, BsmtFullBath, and Condition2 with the subcategory PosN, indicates their strong association with the house sales price.

The incorporation of these key variables in the rebuilt model enhances the accuracy and predictive power of the analysis. By considering the specific subcategories within the Neighborhood and Condition2 variables, as well as the presence of basement full bathrooms, the model can better capture the nuances and complexities of the housing market.

In summary, by converting year value data columns into categorical values and rebuilding the model, the analysis identified five key variables – Neighborhood (Somerset, NridgHt, Crawford), BsmtFullBath, and Condition2 (PosN) – as significant predictors of the house sales price. These variables provide valuable insights into the factors that drive the variation in housing prices and contribute to a more accurate and comprehensive modeling approach.

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:

Before building the model, it is crucial to conduct exploratory data analysis (EDA) to understand the data and its relationship with the dependent variable. This step helps gain insights into the data's behavior, identify patterns, and uncover any potential outliers or anomalies.

To ensure that the model is not overfitting and to eliminate generalized errors, model validation techniques are employed. The objective of model validation is to assess the model's performance on future data by comparing its performance on the training dataset.

Various approaches can be used to prevent overfitting and validate the model. One common approach involves splitting the data into three sets: the training set, the validation set, and the test set. The training set is used to train the model and fit it to the data, while the validation set is used to measure the model's performance and identify the most suitable model. The test set is then used to evaluate the selected model and ensure that the model selection process does not lead to overfitting on the first two datasets.

Another approach to model validation is using cross-validation and the bootstrap method. Cross-validation involves using the test dataset to estimate the model's generalized performance. This method helps evaluate how well the model performs on unseen data and provides a more robust assessment of the model's predictive capabilities.

By employing these validation techniques, we can assess the model's performance, identify any potential issues such as overfitting, and make informed decisions about model selection and generalization. It ensures that the model is reliable and can effectively predict outcomes on new data.

In summary, conducting thorough EDA, implementing data splitting with train, validation, and test sets, and employing techniques like cross-validation and bootstrap validation are essential steps in model validation. These approaches help prevent overfitting, assess the model's performance on unseen data, and ensure the model's generalization capabilities.