Multiple Linear Regression

**Assignment Task:**

Your task is to perform a multiple linear regression analysis to predict the price of Toyota corolla based on the given attributes.

**Dataset Description:**

The dataset consists of the following variables:

Age: Age in years

KM: Accumulated Kilometers on odometer

FuelType: Fuel Type (Petrol, Diesel, CNG)

HP: Horse Power

Automatic: Automatic ( (Yes=1, No=0)

CC: Cylinder Volume in cubic centimeters

Doors: Number of doors

Weight: Weight in Kilograms

Quarterly\_Tax:

Price: Offer Price in EUROs

**Tasks:**

1.Perform exploratory data analysis (EDA) to gain insights into the dataset. Provide visualizations and summary statistics of the variables. Preprocess the data to apply the MLR.

2.Split the dataset into training and testing sets (e.g., 80% training, 20% testing).

3.Build a multiple linear regression model using the training dataset. Interpret the coefficients of the model. Build minimum of 3 different models.

4.Evaluate the performance of the model using appropriate evaluation metrics on the testing dataset.

5.Apply Lasso and Ridge methods on the model.

**Interview Questions:**

1.What is Normalization & Standardization and how is it helpful?

Ans: 1. Normalization:

- Definition: Normalization is the process of scaling numerical data to a standard range, usually between 0 and 1.

- Purpose: It helps to bring all the features to a similar scale, which can improve the performance of some machine learning algorithms that are sensitive to the scale of the input features.

- Formula: The formula for normalization is:

=

where, x is the original value, is the minimum value in the dataset, and is the maximum value in the dataset.

2. Standardization:

- Definition: Standardization (or Z-score normalization) is the process of scaling numerical data so that it has a mean of 0 and a standard deviation of 1.

- Purpose: It makes the data distribution centred around 0 and with a standard deviation of 1, which can be useful for some algorithms that assume normally distributed data.

- Formula: The formula for standardization is:

=

where x is the original value, µ is the mean of the dataset, and is the standard deviation of the dataset.

In summary, normalization scales the data to a range of 0 to 1, while standardization scales the data to have a mean of 0 and a standard deviation of 1.

2.What techniques can be used to address multicollinearity in multiple linear regression?

Ans: Multicollinearity occurs when independent variables in a multiple regression model are highly correlated with each other. This can cause issues with the estimation of the regression coefficients and can make the model unstable. Here are some techniques to address multicollinearity:

1. Feature Selection: Remove one of the highly correlated variables. Choose the variable that is less important to the model or has less theoretical relevance.

2. Principal Component Analysis (PCA): Use PCA to create new uncorrelated variables (principal components) that capture most of the variance in the original data. These components can then be used as predictors in the regression model.

3. Variance Inflation Factor (VIF): Calculate the VIF for each variable to measure the degree of multicollinearity. VIF values above a certain threshold (usually 5 or 10) indicate a problematic level of multicollinearity.

4. Ridge Regression: Ridge regression adds a penalty term to the regression coefficients, which can help reduce the impact of multicollinearity.

5. Lasso Regression: Similar to ridge regression, lasso regression adds a penalty term to the regression coefficients. It has the additional property of performing variable selection, which can help in dealing with multicollinearity.

6. Partial Least Squares (PLS): PLS is a dimensionality reduction technique that aims to find the latent variables that explain the maximum variance in both the predictors and the response variable. It can be used to deal with multicollinearity.

7. Drop One Variable: In cases where two or more variables are highly correlated, you can choose to drop one of them from the model.

It's important to note that multicollinearity does not affect the prediction accuracy of the model, but it can affect the interpretation of the coefficients. Addressing multicollinearity can help improve the stability and interpretability of the regression model.

Ensure to properly comment your code and provide explanations for your analysis.

Include any assumptions made during the analysis and discuss their implications.