Question: From your analysis of categorical variables from the dataset, what could you infer about their effect on target variable?

Answer: Based on analysis, we can see categorical variables have an effect on the target variable (cnt).

1: Year: We can see that bikes in 2019 were high on the road compared to 2018.

2: Month: Their is continuing up trajectory from January to June, and then their down on count of bikes in July, and then it has up trajectory and further from November onwards.

3:Holiday: People use more bikes on holiday.

4: Weekdays: People use more bikes on Monday and Friday.

5: Workingday: People use more bikes on non-working days.

6: Weatersit: People use more bikes when the weather is clean, followed by misty weather.

Question: Why is it important to use drop_first=True during dummy variable creation ?

Answer: Using drop_first=True is a good practice to avoid multicollinearity and ensure the proper functioning of regression models when dealing with categorical variables.

Question: Looking at the pair plot among the numerical variables which one has the highest correlation with the target variable ?

Answer: temp variable has the highest correlation with the target variable

Question: How did you validate the assumption of Linear Regression after building the model on the training set ?

Answer: Here are common ways to validate the assumptions after building the model on the training set:

Residual Analysis:

- Check for Linearity: Plot the residuals against the predicted values. The residuals should be evenly spread around zero, indicating that the relationship between the independent and dependent variables is linear.

Normality of Residuals:

- Examine a histogram or a Q-Q plot of the residuals to check for normal distribution. A normal distribution of residuals is important for valid statistical inferences.

Cross-Validation:

- Perform cross-validation on the training set to assess the model's generalization performance on new, unseen data.

Question: Based on final model , which are top 3 feature contributing significantly towards explaining the demand of the shared bikes

Answer:

1: yr_2019 (Year 2019):

- Coefficient (coef): 998.5904
- t-statistic: 28.596
- p-value: 0.000
- Interpretation: The year 2019 has a substantial positive impact on bike demand, as indicated by the high positive coefficient and low p-value.

2:temp (Temperature):

- Coefficient (coef): 1113.4296
- t-statistic: 22.550
- p-value: 0.000
- Interpretation: Temperature has a significant positive influence on bike demand. As temperature increases, the demand for shared bikes tends to rise.

3: weathersit_Light Snow:

- Coefficient (coef): -357.4601
- t-statistic: -9.317
- p-value: 0.000
- Interpretation: The presence of light snow in the weather has a substantial negative impact on bike demand. In snowy conditions, people are less likely to use shared bikes.

These three features, based on their coefficients, t-statistics, and p-values, are identified as the top contributors to explaining the demand for shared bikes in the final model.

Question: Explain the linear regression algorithm in details ?

Answer:

Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the linear relationship that best predicts the target variable based on the given features. Here's a detailed explanation of the linear regression algorithm:

Simple Linear Regression:

Model Representation:

In simple linear regression, there is only one independent variable (feature).

Multiple Linear Regression:

Model Representation:

In multiple linear regression, there are multiple independent variables (features).

Key Concepts:

- **Gradient Descent:**
  - An optimisation algorithm used to find the minimum of the cost function.
  - It adjusts the parameters in the direction of the steepest decrease in the cost.
- **Feature Scaling:**
  - Normalising or standardizing features to ensure faster convergence in gradient descent.
- **Model Evaluation:**
  - Metrics like R-squared, Mean Squared Error (MSE), etc., are used to evaluate the performance of the model.

Linear regression is widely used for predicting numerical values and understanding the relationships between variables

Question: What is Pearson's R?

Answer:

Pearson's correlation coefficient, often denoted as *r*, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1,

Question: What is scaling ? Why is scaling performed ? What is difference between normalised scaling and standardised scaling ?

Answer:

Scaling:

Scaling is the process of transforming variables to a specific range or standardizing them to have a mean of 0 and a standard deviation of 1. It is performed to ensure that variables are on a comparable scale, especially in machine learning algorithms where the scale of variables can affect the performance of models.

Why Scaling is Performed:

Avoidance of Dominance: Variables with larger scales can dominate those with smaller scales, affecting the behavior of certain algorithms (e.g., gradient descent in linear regression).

Convergence in Optimization: Algorithms that involve optimization, like gradient descent, converge faster when variables are on a similar scale.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

- **Range:** Scales the values between 0 and 1.
- **Advantages:** Preserves the original distribution and is suitable when the data does not have outliers.

Standardized Scaling (Z-score Scaling):

- **Range:** Scales the values to have a mean of 0 and a standard deviation of 1.
- **Advantages:** Useful when the data has outliers and ensures that the scaled data follows a normal distribution.

Question: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

The formula for VIF is:

VIFi =1/1−*Ri^2*

Here, *Ri*2 represents the *R*2 value obtained by regressing the i-th independent variable against all other independent variables. If R^2 is exactly equal to 1, indicating a perfect linear relationship, then the denominator in the VIF formula becomes zero, resulting in an infinite VIF. i.e. Perfect multicollinearity

Perfect multicollinearity can arise due to various reasons, such as:

Redundant Variables: When one variable can be perfectly predicted by a combination of other variables in the model.

Linear Dependence: When there is a linear relationship among the independent variables.