Assignment-based Subjective Questions

Q 1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans-

Boom Bike demand doesn't change whether day is working day or not.
Boom Bike demand in year 2019 is higher as compared to 2018.
Boom Bike demand in the fall is the highest.
Boom Bike demand takes a dip in spring.
Boom Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
The demand of Boom Bike is almost similar throughout the weekdays.

2- Why is it important to use drop_first=True during dummy variable creation?

Ans -

It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.
For Example: We have rows: dteday, unformatted data format it help to clean it make data consistent

3 - **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans-

atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables

4- **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans-

Normality of the error distribution (Normal distribution of error terms). Constant variance of the errors or Homoscedasticity. Less Multi-collinearity between features (Low VIF)

5 - **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Ans-

temp, sep month , rain fall season

General Subjective Questions

Explain the linear regression algorithm in detail.

Basics of Linear Regression

Linear regression assumes that there is a linear relationship between the dependent variable YYY and the independent variable(s) XXX. This relationship can be represented by the equation:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

where:

- YYY is the dependent variable.
- X1,X2,…,XnX_1, X_2, \ldots, X_nX1,X2,…,Xn are the independent variables.
- β0\beta_0β0 is the y-intercept (constant term).
- β1,β2,…,βn\beta_1, \beta_2, \ldots, \beta_nβ1,β2,…,βn are the coefficients that represent the change in YYY for a one-unit change in the corresponding XXX.
- ϵ\epsilonϵ is the error term (residual) that captures the deviations of the actual values from the predicted values.

2. Types of Linear Regression

- **Simple Linear Regression:** Involves one independent variable. Y=β0+β1X+ϵY = \beta_0 + \beta_1X + \epsilonY=β0+β1X+ϵ
- **Multiple Linear Regression:** Involves multiple independent variables. Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

3. Assumptions of Linear Regression

1. **Linearity:** The relationship between the dependent and independent variables is linear.
2. **Independence:** Observations are independent of each other.
3. **Homoscedasticity:** The residuals (errors) have constant variance at every level of XXX.
4. **Normality:** The residuals of the model are normally distributed.
5. **No Multicollinearity:** Independent variables are not highly correlated with each other.

Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. This collection was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show that statistical properties do not provide a complete picture of the data. Each dataset consists of eleven (x, y) points.

Characteristics of Anscombe's Quartet

Each of the four datasets in Anscombe's quartet has the following properties:

- The mean of the x values is 9.
- The sample variance of the x values is 11.
- The mean of the y values is approximately 7.50.
- The sample variance of the y values is approximately 4.12.
- The correlation between x and y is approximately 0.816.
- The linear regression line is y=3+0.5xy = 3 + 0.5xy=3+0.5x.
- The coefficient of determination (R^2) is 0.67 for the linear regression.

What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which two variables are linearly related. The value of Pearson's R ranges from -1 to 1, where:

- r=1r = 1r=1 indicates a perfect positive linear relationship,
- r=−1r = -1r=−1 indicates a perfect negative linear relationship,
- r=0r = 0r=0 indicates no linear relationship.

Formula

The Pearson correlation coefficient for two variables XXX and YYY is calculated as:

r=∑i=1n(Xi−X¯)(Yi−Y¯)∑i=1n(Xi−X¯)2∑i=1n(Yi−Y¯)2r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}r=∑i=1n(Xi−X¯)2∑i=1n(Yi−Y¯)2∑i=1n(Xi−X¯)(Yi−Y¯)

where:

- XiX_iXi and YiY_iYi are the individual data points,
- X¯\bar{X}X¯ and Y¯\bar{Y}Y¯ are the means of XXX and YYY respectively,
- nnn is the number of data points.

Interpretation

- **1**: A perfect positive correlation; as XXX increases, YYY increases linearly.
- **-1**: A perfect negative correlation; as XXX increases, YYY decreases linearly.
- **0**: No linear correlation; changes in XXX do not predict changes in YYY.

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling

Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range of the features in the data. The primary goal of scaling is to ensure that different features contribute equally to the analysis and to improve the performance of machine learning algorithms. Some algorithms are sensitive to the scale of the data, and scaling can help achieve better convergence and results.

Why is Scaling Performed?

1. **Equal Contribution of Features**: Features with larger ranges can dominate the model training process, leading to biased results. Scaling ensures all features contribute equally.
2. **Improved Model Performance**: Many machine learning algorithms (e.g., gradient descent-based algorithms, k-nearest neighbors, SVMs) perform better and converge faster when the data is scaled.
3. **Reduction of Numerical Instability**: Scaling can help reduce numerical instability and improve the precision of calculations, especially for algorithms that involve matrix operations.

Types of Scaling

1. Normalized Scaling

Normalization, also known as min-max scaling, rescales the features to a fixed range, usually [0, 1] or [-1, 1]. The formula for normalization is:

Xnorm=X−XminXmax−XminX_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xnorm=Xmax−XminX−Xmin

where:

- XXX is the original value.
- XminX_{\text{min}}Xmin and XmaxX_{\text{max}}Xmax are the minimum and maximum values of the feature, respectively.

When to Use Normalization:

- When you know that the data has a bounded range.
- When the distribution of the data is not Gaussian.
- When you want the data to be on the same scale without distorting differences in the ranges.

Advantages:

- Preserves the relationships between the original data points.

- Useful for algorithms that do not assume any particular distribution of the data (e.g., k-nearest neighbors).

Disadvantages:

- Sensitive to outliers, as they can significantly affect the min and max values.