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?

From the analysis of the dataset containing six categorical variables—season, year (yr), month (mnth), holiday, weekday, and weathersit—the following insights can be inferred:

1. **Season and Active Customers**: Fall (Autumn) appears to attract the maximum number of active customers, with September being the peak month within this season.
2. **Yearly Sales Comparison**: The year 2019 recorded higher sales compared to 2018.
3. **Impact of Holidays**: Holidays negatively impact the count of active users, resulting in a decrease.
4. **Weather Conditions and User Activity**:
    - During heavy rain (weathersit indicating heavy rain), there are no users observed.
    - Conversely, partly cloudy or clear sky conditions (weathersit indicating clear or partly cloudy) show the highest count of active users.

These observations highlight the significant influence of seasonal variations, yearly trends, holiday periods, and weather conditions on user activity, providing valuable insights for further analysis and decision-making.

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

Not using drop_first=True when creating dummy variables can lead to multicollinearity issues where the dummy variables become correlated with each other. This correlation can introduce redundancy in the model, which is undesirable for our analysis. Therefore, setting drop_first=True ensures that one dummy variable is dropped to mitigate multicollinearity, maintaining the integrity and interpretability of the regression model.

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

atemp and *temp* has the highest correlation with the target variable.

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

One fundamental assumption of a linear regression model is that the error terms should follow a normal distribution when plotted on a histogram. In our analysis, we observed that the error terms indeed conform to this normal curve. This validates the assumption of normality for the error terms in our linear regression model.

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

Ans. The top 3 features are: Temp, Year (positively influencing) and hum (negatively influencing).

General subjective questions

Q 1: Explain the linear regression algorithm in detail.

Linear regression is an interpolation technique used to predict the correlation between variables and understand how an independent variable is influenced by dependent variables. Once data has been thoroughly examined and cleaned through exploratory data analysis, it is split into a training set for model training and a testing set to evaluate model performance against actual outputs.

During model development, variables' collinearity is assessed, and only relevant variables are used for training. The model's quality is evaluated using metrics such as the R-squared value and p-values of dependent variables. Iterative steps like feature elimination may be employed to refine the model further.

Linear regression assumes that the error terms follow a normal distribution. After confirming the model meets this assumption, it is tested using the test dataset. Insights and predictions drawn from the final model can then be applied to new data points within the model's scope, providing valuable insights into the relationships and predictions of interest.

Q 2 : Explain the Anscombe’s quartet in detail.

A regression model isn't always exact and can be influenced by cleverly constructed data. In some cases, different datasets can appear identical to a regression model after training, despite having distinct characteristics. One famous example demonstrating this phenomenon is Anscombe's quartet, consisting of four datasets with identical descriptive statistics but unique patterns and relationships.

Q 3: What is Pearson’s R?

Pearson's correlation coefficient, also known as Pearson's R, quantifies the strength of the linear relationship between two variables. It is extensively used in linear regression analysis. The coefficient ranges from -1 to +1: a value of +1 indicates a perfect positive linear correlation, while -1 indicates a perfect negative linear correlation. Values between -1 and +1 denote the degree of linear association between the variables, with closer values to these extremes indicating stronger correlation.

Q 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is essential for ensuring a model operates effectively within a suitable range of coefficients. For instance, consider two independent variables, "price" and "months," influencing car sales. The price variable typically spans a much larger range compared to months, which only ranges from 1 to 12. Scaling

the price variable appropriately prevents numerical instability in the model. There are two primary types of scaling:

1. **Normalized Scaling**: This type of scaling transforms the data distribution into a Gaussian (normal) distribution without imposing a specific range. It is commonly used in neural networks.
2. **Standardized Scaling**: In the example mentioned earlier, standardized scaling is employed. Here, variable values are adjusted to fit within a defined range, ensuring consistency and preventing large discrepancies in scale between different variables in the model.

Q 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

When the dependent variable and independent variable(s) exhibit perfect correlation, the R-squared value reaches 1. Consequently, the VIF (Variance Inflation Factor), calculated as 11−R2\frac{1}{1 - R^2}1−R21, tends towards infinity in such cases.

Q 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q plot is a graphical tool used to evaluate whether datasets follow the same statistical distribution. It is especially valuable in linear regression, where distinct testing and training datasets are provided. In such cases, it is crucial to verify if both datasets originate from the same underlying distribution to ensure the model's reliability.