Linear Regression Assignment Subjective Questions

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- In categorical variables plot of '*weekday'* and '*workingday'* we can observe the medians of all categories are not having much difference, which could mean that change in these variables is not creating much impact on target variable
- By observing the boxplot of feature '*holiday'* we can say that people are using shared bikes mostly in non holidays may be for travelling to work.
- By observing boxplot of '*weathersit*' we can say that people are more interested to use bike sharing on clear days comparatively when there is '*mist'* or '*lightsnow'*
- Even the same can be observed with feature variable '*season'* that people are most likely to use bike sharing platform in '*fall'* and '*summer'* rather than in '*spring'* or '*winter'*
- We can observe there is raise in year on year usage of bike sharing platform.

Q2. Why is it important to use drop_first=True during dummy variable creation?

- While converting categorical variables to dummy variables we can answer the availability of one variable with absence of other variables, this avoids model getting into dummy variable trap.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- Features *'atemp'* and *'temp'* are having highest and near to same correlating with target variable *'cnt'*.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- We validate the assumption of error terms linearly distributed by plotting distribution plot to residuals.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Based on the model top 5 features significantly impacting:
    - *'temp'* is highly correlated with target variable
    - *'hum' & 'windspeed'* are negatively correlated highly with target variable
    - *'season' and 'weathersit'* are other 2 features which are contributing significantly.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

- Linear Regression algorithm is way of predicting continuous target variable based on available features or variables, where we can find some linearity or correlation between input and target variables.
- Linear regression algorithm tries to fit a line in the data using least squares and trains model.
- This is considered as supervised learning model, where we divide existing data set to train and test data, we train the model using train data and will try to predict target using test data.

Q2. Explain the Anscombe's quartet in detail.

- *Anscombe's Quartet* is the model example to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

Q3. What is Pearson's R?

- The Pearson's R is a correlation coefficient that measures linear correlation between 2 sets of data.
- It is the ratio between the covariance of two variables and the product of their standard deviations
- This is the number between -1 and 1 that measures the strength and direction of the relationship between two variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is converting numerical data to smaller values or into specific range of data for easy understanding and processing.
- Data may contain features highly varying in magnitude , units and range. If scaling is not done model takes only magnitude into account and not unit, which is incorrect modeling. To solve this we will do scaling to bring all variables to same ranges.
- Standardization scaling replaces all the values by their z score, which will have mean and standard deviation
- Min Max scaling brings all the data into the range of 0 to 1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- If the correlation between all input variables other than target variables is perfectly aligned or correlated then the VIF value will become infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q-Q plot helps us to indentify the if the data has the same kind of distribution.
- It helps to determine if two data sets come from populations with a common distribution.