Bike Sharing Assignment

Assignment-based Subjective Questions &
General Subjective Questions

Name: Arunkumar

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

   - **People are more likely to use Bike sharing in summer & fall season when compared to winter & spring**
   - **More number of peoples used bike sharing 2019 when compared to 2018**
   - **More number of peoples used bike share in mid months of the year - June, July, August**
   - **More number of people used bike on weather situation such as low cloud or mid cloud when compared when compared the light rain situation. No one used bike when there was heavy rain**

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   It's important to get 'm-1' dummies out of 'm' categorical levels by removing the first level to reducing the extra column created during dummy variable creation

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   'temp' has the highest correlation with the target variable among the numerical variables (not mentioning 'atemp' because the correlation of temp & atemp is 0.99)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   After building the model, I would look at the Adjacent R2 and R2 to see if it's high or low to know whether the model fitted properly. Then, I'll check Prob (F-statistic), P-value & VIF. In the end, I'll validate the model using the test data and check the R2

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

   'temp', 'yr', 'light_rain'

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression is ML algorithm used to perform regression task. We will create Regression model using the Independent variable to predict the Dependent variable on the train data first. Then we will the model to predict the dependent variable on the Unseen data. Simply put, to find the pattern in data and model it to predict on the unseen data

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four data sets that have similar statistics, but they have completely different distributions and look different when graphed. Each dataset consists of 11 points

3. What is Pearson's R?

Pearson's R is a measure of linear correlation between two sets of data. It's a normalised measurement of the covariance, such that the result always has a value between –1 and 1

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

Scaling is used to standardise the independent features present in the data in a fixed range. We do scaling to pre-process the different magnitudes among different

independent variables. Normalized scaling rescales the feature values between 0 and 1. Standardization rescales the feature values with 0 mean value and variance equals to 1

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

When there is perfect correlation between two independent variables, we get R2 = 1 which leads to 1/(1-R2) infinity

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

Q-Q plot is used plotting two sets of quantiles against one another. In linear regression, its used for comparing two probability distribution by plotting their quantiles against each other.

Thank you

Arunkumar