Linear Regression Assignment

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?

- I visualized the relationship between the categorical variables and the target variable. It was seen that during the weather situation that Clear, few clouds, partly cloudy, a high number of bike rentals were made and with the median being 50,000 approximately. Similarly, certain inferences could be made 'season' and 'yr' as well. Also, during model building on inclusion of categorical features such as yr, season, etc. We saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset

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

- During dummy variable creation, it is best to use drop_first=True. Otherwise, we will get a redundant feature. We need to map spring, summer, fall and winter to respective seasons, so that these column values will be further used to generate dummy variables as it is a Categorical Nominal Type Data.

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

- 'registered' has the highest correlation with the target variable 'cnt'

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

- By drawing the distribution of residuals against levels of the dependent variable. One can use a QQ-plot and measure the divergence of the residuals from a normal distribution.

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

-

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a linear model. The model that assumes the linear relationship between input variables (x) and single output variable (y). More specifically, that 'y' can be calculated from the linear combination of input variables (x).

When we have a single input variable 'x' , the procedure is referred as a simple linear regression. When we have multiple input variables, literature from statistics refers to the method as multiple linear regression.

Different techniques can be applied to prepare and train the linear regression equation from data, some common way of which is called Ordinary Least Squares. The simple linear model would be represented by y = B0 + B1*x

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet states that the importance of a plotting data to confirm that a validity of the model fit. In a single panel, a Pearson correlation between the variable 'x' and 'y' values are same, r = . 816. and also the 4 different data sets are also same in terms of the mean and variance value of the variable 'x' and 'y' values.

3. What is Pearson's R?

- The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

It is a bivariate statistical model used to analyzes 2 variables. Pearson's correlation is used to test the associative research hypothesis as long as variables being analyzed are both quantitative.

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

- Scaling is the procedure of measuring and assigning the objects to the numbers

according to the specified rules. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

In normalisation Minimum and maximum value of features are used for scaling

Mean and standard deviation is used for scaling.

Normalisation is used when features are of different scales.

Standardisation is used when we want to ensure zero mean and unit standard deviation.

Normalisation Scales values between [0, 1] or [-1, 1].

Standardisation is not bounded to a certain range.

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

- If there is a perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

- The quantile-quantile (q-q) plot is a graphical technique to determine 2 data sets

come from populations with a common distribution.

Q-Q plot used to assess if the set of data pausibly came from the theoretical distribution Normal, exponential or Uniform distribution