Linear Regression Assignment – Santosh Kumar Ramarathnam

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)
    a. Weather factors such as temperature, weather situation, humidity had a strong effect on the dependent variable (cnt)
    b. Seasonality – spring demand seems counter intuitive. Summer and fall have a strong positive effect on the bike rentals (cnt).
    c. 2019 shows significantly higher demand than 2018.
    d. May to October is the busiest season.
    e. The last 4 days of a week have slightly higher rentals than the first 3 days (Monday through Sunday order). This behavior is more pronounced in 2019 than in 2018. 2018 had even spread across the days of the week.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
    a. It reduces the number of columns when projecting a categorical variable and merging it with the original dataset.
    b. When the value of all other columns (dummy vars generated) is zero, this can be identified.
    c. This is done to reduce multi-collinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    a. feelslike_t (atemp) and actual_t(temp) are both highly correlated with the target variable (cnt).
    b. In the training data subset, their correlation was 0.65 against cnt.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. By performing residual analysis and ensuring the normality of error terms existed.
    b. The error terms are independent of each other and have constant variance.
    c. There is linear relationship between the predictor variables and the target variable.
    d. Low p-values for the features indicate that we can reject the null hypothesis that the corresponding feature does not impact the target variable. In our model summary calculations, we noticed that all of the features had p-values less than 0.05 and therefore conclude that the features are significant.
    e. High R-squared values 80+% means that the model could explain 80% of the variation in the actual outcome (rentals).

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. feelslike_t
    b. weathersit
    c. windspeed.

Linear Regression Assignment – Santosh Kumar Ramarathnam

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    - Linear regression is an ML algo used in supervised learning.
    - It performs the task of predicting the value of a target variable given a set of independent variables. It identifies a linear relationship and fits a line between the independent variables and the target variable.
    - There are two types of linear regression:
        - Simple Linear Regression – which takes the form y = mx + c
        - Multiple Linear Regression which takes the form y = m1x1 + m2x2 + … + c
2. Explain Anscombe's quartet in detail. (3 marks)
    a. Anscombe's quartet is a set of 4 datasets that were particularly created to emphasize the importance of data visualization even before doing model fitment and subsequent analysis.
    b. These 4 datasets fool the linear regression model where in reality datasets contain non-linear relationships and outliers, which aren't handled by the model.
    c. The statistical information such as mean, sample size, SD about these data sets are similar in nature however they generate completely different plots.

3. What is Pearson's R (3 marks)
    a. Pearson's R is a measure of linear correlation between two sets of data. It is known by many names: PPMCC, correlation coefficient, PCC.
    b. For a given pair of random variables a,b the Pearson's R is calculated as covariance of a & b divided by the product of their standard deviation.
    c. It ranges from -1 to 1, with an absolute value of 1 implying a perfect linear relationship (as a increases b increases) and -1 implying the vice versa (as a increases b decreases or vice versa).

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    a. Scaling is a data preparation step.
    b. It is done to normalize the data within a range, and if not done, can lead to incorrect modeling since magnitude would be taken into consideration.
    c. Normalized scaling is min-max scaling to fit the data between 0 and 1.
    d. Standardized scaling replaces the values with their Z-scores.
    e. Normalization loses some precision in the data over standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
    a. It happens because of multi-collinearity.
    b. VIF = inf shows that the variable is perfectly represented by another variable within the data set and can be eliminated in the feature selection process.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

a. Q-Q plot is a quantile-quantile plot – it is a probability plot to compare two probability distributions by plotting their quantiles against each other.
b. It is useful in linear regression cases where test and train data are obtained from different sources and we want to identify or confirm that they came from populations with the same distribution.