LINEAR REGRESSION ASSIGNMENT

Paramesh E

- 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?

1. 1. season Boxplot shows that during fall season maximum bikes are rented in plot 1
2. 2. year Boxplot shows that, in 2019 most bike as rented
3. 3. comparision between holiday, working day and weekday, the more bikes are rented in weekdays rather than weekend and holiday
4. 4. weatherit Boxplot shows that the most bikes are rented when atmospher is Clear, Few clouds, Partly cloudy
5. 5. month Boxplot shows that, in the month of september the maximum bikes are rented and followed by august and november month

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

- It reduces the extra columns created, According to (n-1)Rule while creating the dummy variables to reduce the complexicity we have to create (n-1)columns, Hence it reduces the correlations created among dummy variables.

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

- based on pair-plot among the numerical variables 'temp' has the highest positive correlation 0.63 with the target variable 'cnt'

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

- There should be a linear and additive relationship between dependent variable and predictor variable.
- There should be no correlation in between the residual terms. independent variables should not be correlated. I.e multicollinearity.
- The error terms must have constant variance known as homoscedasticity. The non-constant presence known as heteroscedasticity.
- its known to be error terms are normally distributed

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

- Temperature (Temp) : A coefficient value of ‘0.564438’ indicated that a temperature has significant impact on bike rentals
- Light Rain and Misty (weatherit) : A coefficient value of ‘-0.307082’ indicated that the light snow and rain deters people from renting out bikes
- Year (yr) : A coefficient value of ‘0.230252’ indicated that a year wise the rental numbers are increasing

- General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable (target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression..
- **Advantages:**
- Linear Regression is simple to implement and simple in usage.
- As compared to other algorithms it has Less complexity.
- Linear Regression may lead to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization techniques, and cross-validation.
- **Drawbacks:**
- In this algorithm Outliers affects badly.
- It over-simplifies real-world problems by assuming a linear relationship among the variables, hence not recommended for practical use-cases.

2. Explain the Anscombe's quartet in detail.

- We often look for summary statistics during EDA (Exploratory Data Analysis). But, sometimes these statistics may give us wrong interpretation of the data. In 1973, a statistician Francis Anscombe demonstrated it with the help of four datasets known as Anscombe's quartet.
- All the datasets have the same statistical summary: mean, standard deviation, same correlation between x and y
- datasets are so much different while they seemed the same by looking at the statistical summary. Now, we realize the importance of graphing data before analyzing it.

3. What is Pearson's R?

- Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

- Scalling is setting whole this in some certain ranges
- Scalling performed because make easier for processing and can unerstand very well
- The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 is also known as unit variance.

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

- VIF shows infinity because there is perfect correlation in between two variables and In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop any one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable can be expressed exactly by a linear combination of other variables

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

- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us access if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
- This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.