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?

Answer:

T (temp) - A coefficient value of '0.3204' indicated that a unit increase in temp variable increases the bike hire numbers

Windspeed - A coefficient value of '-0.1192' indicated that windspeed unit increase the bike hire numbers units.

Year (yr) - A coefficient value of '0.2353' indicated that a unit increase in yr variable increases the bike hire numbers

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

Annswer:

Whenever we are creating the dummy variables this command drop_first=True will help us to reduce the extra columns created when we creating the dummy variables we use the drop first to drop all the first variables. This could reduce the error while creating

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

Answer:

The variable temp having highest correlation with 0.3204 co-efficient based on the final model Interpretation based on heat map the highest correlation found between cnt and temp.

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

Answer:

The residual terms are pretty much normally distributed for the number of test points we took. Remember the central limit theorem which says that as the sample size increases the distribution tends to be normal. But there is a change in perfect curve which would say its not an perfect distribution

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

Answer:

'const', 'temp', 'yr' variables having significant contribution towards the demand of the boom bike sharing system.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is statistical regression method which mainly used to show the relationship between independent variable and dependent variable .If there is an single input variable such linear regression is called simple linear regression. If there are more than one input variable the its called multiple linear regression. The general formula for linear regression is Y=mx+b

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson's R?

Answer:

The Pearson's Correlation Coefficient(Pearson's r),It is like statistic that measures the linear regression between two variables. Like all the correlations it also finds the correlation between -1.0 and +1.0 but it cannot find the difference between dependent and independent variables.basically Pearson's r is covariance of two variables between standard deviation.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within particular range basically scaling is most significant steps that follows in all the linear regression methods
Why Scaling
All the collected data set will have either high or low units or range. If scaling is perform it will bring all the variables to the same level of units if its not done it will take the variables as it is showing in the original data set that may not give correct model.
Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Min-Max Scaling: x=x-min(x)/max(x)-min(x)

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

Standardization:x=x-mean(x)/sd(x)

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If VIF = Infinty this will give perfect correlation between independent variables. In case of perfect correlation we get R2 = 1, which lead to 1/(1-R2) infity. We have to drop one of the variable from dataset

An infinite VIF value indicates that the corresponding variable may 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

Q-Q Plots (Quantile-Quantile plots) is a graphical tool that help us to assess the dataset basically Q-Q Plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. It helps to determine if two data sets come from populations with a common distribution. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line.