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-:

cnt is the dependent variable

Season, yr, mnth, workingday, weathersit, weekday, holiday are the categorical variables. The inference about their effect on the dependent variable is stated below.

Season - count of total rental bikes (cnt) is least in Spring and highest in Fall season.

yr - count of total rental bikes (cnt) has increased in the year 2019 from 2018.

mnth - count of total rental bikes (cnt) is minimum in the month of January and maximum in the month of September

workingday - count of total rental bikes(cnt) is higher on non-working days

weathersit - count of total rental bikes(cnt) is higher for the Clear, Few clouds, Partly cloudy, Partly cloudy weather and zero for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.

weekday - Average count of total rental bikes (cnt) is higher in weekdays

holiday - count of total rental bikes (cnt) is lesser on holidays.

---

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

Drop_first = True is important because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. And we drop one variable while creating dummy variable is because the nth variable value can be found by n-1 variable values.

---

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

Temp 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?

- For training set Error terms are normally distributed and has mean value around zero which satisfies the assumptions about residuals of linear regression.

- There is no multicollinearity in the training data as the parameter which has high VIF has been dropped while training the model. Therefore assumptions about estimator satisfies.

- As there is a linear relationship between dependent and independent variables, therefore Linearity Assumption satisfies on

training set.

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

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are temp, season_Winter and month_Sep and there is increment seen in count of rental bikes from 2018 to 2019.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised Machine Learning Algorithm. It is used for linear relationship between the dependent and independent variables.

Explains change in dependent variable and change in the values of predictors. In simple linear regression one variable changes at a time.

In multiple linear regression multiple variable changes at a time.

Regression guarantees interpolation of data and shows correlation not causation.

Data allows fixed parameters means finite number of parameters.

Linear regression is a method of finding best straight line fitting to the given data. Shortcomings of Linear regression is that they are sensitive to outliers. Models linear relationship only and assumptions are required to make inference.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, r = . 816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.

3. What is Pearson's R?

If we keep on increasing power of R, the Pearson's R value drops.

It is a measure of linear correlation between two sets of data.

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

Scaling is a step of pre-processing and is applied to independent variables to normalize the data in a particular range.

Scaling is done to bring all the variables to the same level of magnitude.

Normalization or Normalized scaling brings all data in 0 to 1 range. Also known as min max scaling.

Standardized scaling or Standardization replaces the values by their Z scores.

Normalization loses some information in the data, mostly about outliers.

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

When there is perfect correlation between independent variables than value of VIF is equal to infinte.

We know the formula of VIF i.e, VIF = 1/1-Ri2

Whenever there is perfect correlation among independent variables, then R2 = 1

Putting the value of R2 in formula it gives VIF = infinite.

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

Q-Q plot is a quantile quantile plot. These are plots of two quantiles against each other. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties are similar or different in the two distributions.