BOOM BIKE RENTAL ASSIGNMENT

CREATED BY: DR RANJEET SINGH MAHLA
SUBMITTED BY: DR RANJEET SINGH MAHLA

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: From the analysis of categorical variables from the datasets I can infer some information about dependent variables.

- The bike rentals are high in summer and fall of 2019
- Bike rentals are high in September and October of 2019
- Bike rentals are high on Saturday, Wednesday and Thursday of 2019
- Bike Rental are high on clear weather days
- Bike rentals are high on holidays
- More bike rentals in the year 2019 means people are following the idea of boom bike rentals

Q2: Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: “drop_first=True” helps in reduction of creation of extra column created during the dummy variable creation and hence avoid redundancy of any kind.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: “temp” variable has the highest correlation with the target variable.

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the holiday variables.

GENERAL SUBJECTIVE QUESTIONS

Q1: Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear Regression is a machine learning algorithm used for supervised learning. Using linear regression algorithm, we can predict a dependent variable based on independent variables. Dependent variable is also known as target variable. Linear regression establishes a linear relationship between dependent and given independent variables.

There are two types of linear regression:

(1) Simple linear regression: Is used when a single independent variable is used to predicting the value of the target variable.

(2) Multiple linear regression: Is used when multiple independent variables are used to predict the numerical value of the target variable

A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

Q2: Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet consists of four data sets that have nearly the same statistical observation (i.e., identical simple descriptive statistics) but have very different distributions and appear very different when presented graphically. The statistical analysis provides same information on variance and mean for each x and y points in all four datasets. However, when these datasets are plotted, they look different.

Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

Anscombe's quartet let us understand the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can

help you identify the various anomalies present in the data such as diversity of the data, outliers, and linear separability of the data. By the ways linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Q3: What is Pearson's R? (3 marks)

Pearson's R is Pearson corelation coefficient defines the linear relationship between the covariance of two variables and the product of standard deviation between the two sets of data.

Basically, it is the ratio between the covariance of two variables and the product of their standard variations. Essentially it is a normalized measurement of the covariance such that the result always has a value between plus (+) 1 and minus (-) 1. The covariance itself can reflect a linear corelation of variables and ignores other types of relationships and corelations.

It is defined as a covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment of the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

It can be applied to both population and samples

When it is applied to a population, it is represented as Greek letter Rho (ρ) and may be referred as the population correlation coefficient or the population

For a given pair of random variables the formula can be draws as follows

$\rho XY = cov(X, Y)/ \sigma X \sigma Y$

cov =covariance
σX = standard deviation of X
σY = standard deviation of Y

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R2 (squared value) is 1 in this case. Leads VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression. Infinite VIF basically indicates that the corresponding variables may be expressed exactly by a linear combination of other variables which show an infinite VIF.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.

The situation when two distributions are similar the points in the Q-Q plot will lie on the line y =x. The situation when two distributions are linearly related the points on Q-Q plots will lie on a line but not necessarily on y =x. Q- Q plots also can be sued as a graphical means of estimating parameters in a location scale family of distribution.

It is used to compare the shape of distribution which provides a graphical view of how properties such as location, scale and skewness are similar or different in the two distributions.