BIKE SHARING ASSIGNMENT

Linear Regression Assignment

PRIYANKA KUMARI

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)***

   Answer: The Inference that I could derived were:
   1. Most of the bike were book in season Fall whereas least booking in season spring.
   2. Maximum Bike booking were done in the month of May, Jun, July, Aug and sept

      and least booking is done in month Jan.
   3. Most of the booking were done on weather_1 which is Clear, few clouds, partly cloudy, partly cloudy and least booking in weather_3 which is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
   4. Almost 97.6% Bike booking were happening when it is not a holiday which means this data is clearly biased.
   5. Weekday shows very close trend from Thurs to Sun.
   6. No noticeable change in bike demand with working days and non-working days.

2. *Why is it important to use **drop first=True** during dummy variable creation? (2 mark)*

   Answer: drop first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   E.g., Let's say we have 12 types of values in Categorical column Month and we want to create dummy variable for that column. If 1st variable is Jan, 2nd is Feb ,3rd is march and so on till Nov then It is obvious dec. So, we do not need 12th variable to identify the dec. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

   Answer: Looking at the pair-plot among the numerical variables, "temp" has the highest corelation with the target variable "cnt".

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

   Answer:

   1. Checking whether the errors are normally distributed by distplot.

2. Checking whether the linear relation can be seen between target and predictor variable by using pairplot
3. No patterns are observed in the scatter plot of residuals vs fitted value

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

Answer: As per our final Model, the top 3 predictor variables that influences the bike booking

1. Temperature (temp) - A coefficient value of '0.4782' indicated that a unit increase in temp variable increases the bike hire numbers by 0.4782 units
2. Weather_3 - A coefficient value of '-0.2860' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2860 units
3. Year - A coefficient value of '0.2341' indicated that a unit increase in yr. variable increases the bike hire numbers by 0.2341 units.

General Subjective Questions

1. *Explain the linear regression algorithm in detail. (4 marks)*

   Answer: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. It is one of the very basic forms of machine learning where we train a model to predict the behaviour of our data based on some variables. In the case of linear regression as we can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated. Linear regression is used to predict a quantitative response Y from the predictor variable X. The independent variable is known as the predictor variable, and the dependent variables are known as the output variables.

   Linear Regression models classified into types depending upon the no of variables:

   1. simple Linear Regression: This is used when independent variable is one.

      $$Y = \beta_0 + \beta_1X$$

   2. Multiple Linear Regression: This is used when independent variable is more than one.

      $$Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn$$

   Where Y is the target variable and X is the independent Variable. β_0 is the intercept and β_1 is the coefficient of X. when we train the model, we look for the best fit line to predict the value of target variable for the value of X. once we got the best fit model, we get the value of β_0 and β_1.

   The strength of a linear regression model is mainly explained by R^2, were
   R^2 = 1 - (RSS/TSS)

   RSS: Residual sum of squares

   TSS: Total sum of squares

2. *Explain the Anscombe's quartet in detail. (3 marks)*

   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.

   It was built by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance**,** and mean of all x, y points in all four datasets. Data must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3. *What is Pearson's R? (3 marks)*

Answer Correlation coefficients are used to measure how strong a relationship is between two variables. There are so many types of correlation coefficient, but the most popular is Pearson's correlation. It ranges from +1 to -1 whereas 1 indicates a strong positive relationship between the two variables means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. -1 indicates a strong negative relationship between the two variables means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. (almost) perfect. A result of zero indicates no relationship between two variables means that for every increase, there isn't a positive or negative increase. The two just aren't related.

4. *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)*

Answer: Scaling It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Most of the times, data set contains variables highly varying in unit's magnitude and range. there is a chance that higher weightage is given to features with higher magnitude. This will impact the performance of the machine learning algorithm and obviously, we do not want our algorithm to be biased towards one feature.

Normalization: It is a scaling technique in which values are distributed between 0 and 1. It is also known as Min-Max scaling.

Formula for Min-Max Scaling**:** x`=x−min(x)/max(x)−min(x)

Standardized: It is a scaling technique which brings all the data into a standard normal distribution with mean 0 and standard deviation 1.

Formula for standardized scaling: x`=x−mean(x)/standard deviation (x)

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? (3 marks)***

   Answer: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. This happens when there is multicollinearity.

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

   Answer: Q-Q plot is a graphical tool which is used to see if dataset came from normal, exponential or uniform distribution. Also, it determines if two datasets have common distribution.

   Outliers, changes in symmetry, different scales can be identified from this plot. It is used to check if two datasets have same common distribution, have common scale, have same distributional shapes.

   Using Q-Q plot we can have the interpretations:

   1. All point of quantiles lies on or close to straight line from x -axis then It is a similar distribution.
   2. y-quantiles are lower than the x-quantiles.
   3. x-quantiles are lower than the y-quantiles
   4. All point of quantiles lies away from the straight line from x -axis then it is different distribution