Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?

=> Categorical variables have different effects on the dependent variable such as:

1. Year 2019 shows more usage of bikes than year 2018.

2. Season 3 shows most number of bike usage

3. Month 9 has shown most number of bike usage

4. Weathersit 1 has shown to be the most likely weather for bike usage.

Q2. Why is it important to use drop_first=True during dummy variable creation?

=> The use of drop_first = True during creation of dummy variables, drops the first column from the "n" number of columns created during transformation. The dropped column is described by rest of the columns, thus helping in removing unnecessary excess columns.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

=> Among the numerical variables, "temp" and "atemp" show equally high correlation with target variable that is 0.63

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

=> 1. By checking distplot of residual, which came out to be normally distributed.

2. By checking for any patterns in scatterplot between the predicted values and residuals, which did not show any pattern.

3. By checking the value of mean squared error, which came out to be closer to 0.

Thus, by all these parameters I validated the assumptions of Linear Regression.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

=> "temp" , "atemp" and "year" are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

=> Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

Equation:

$y = mx + c$

where,

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best $\theta 1$ and $\theta 2$ values.

c: intercept

m: coefficient of x

Once we find the best c and m values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Q2. Explain the Anscombe's quartet in detail.

=> 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 constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these 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.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features 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

Q3. What is Pearson's R?

=> In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

It is calculated as (x(i)-mean(x))*(y(i)-mean(y)) / ((x(i)-mean(x))2 * (y(i)-mean(y))2

Pearson's correlation coefficient is the 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 about the origin) of the product of the mean-adjusted random variables, hence the modifier product-moment in the name.

Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Q4. 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 a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

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

   Formula: x-xmin/(xmax – xmin)

2. Standardized 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).
   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.
   Formula: x-x.mean()/x.std()

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

=> If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. 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 one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable

may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6. 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 assess 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.

Advantages:

1. It can be used with sample sizes also
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Uses:

1. Check if two datasets come from populations with a common distribution.
2. Check if two datasets have common location and scale
3. Check if two datasets have similar distributional shapes
4. Check if two datasets have similar tail behaviour