Bike Sharing Assignment – Mohammed Suleman

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)**

   The categorical variables examined in the dataset include 'season,' 'weather situation,' 'holiday,' 'month,' 'year,' and 'weekday.' These categorical values were visually represented using boxplots, and they exhibited the following effects on our dependent variable:

   A) Season: The boxplot revealed that the spring season had the lowest 'cnt' values, while the fall season had the highest 'cnt' values. Summer and winter seasons displayed intermediate 'cnt' values.

   B) Weather Situation: Notably, when heavy rain or snow was present, there were no users, indicating that such weather conditions are highly unfavorable for bike rentals. The highest bike rental counts were observed when the weather situation was described as 'Clear, Partly Cloudy.'

   C) Holiday: Rental counts decreased during holidays, suggesting that bike rentals are less popular on holidays compared to regular days.

   D) Month: September witnessed the highest number of rentals, while December had the lowest. This observation aligns with the 'weather situation' analysis, as December typically experiences heavy snowfall, making it less conducive for bike rentals.

   E) Year: In terms of years, there were more bike rentals in 2019 compared to 2018.

These insights provide valuable information about how these categorical variables relate to the dependent variable, 'cnt,' and can inform decision-making in bike rental management.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   If the first column is not omitted, it will lead to a situation where all the dummy variables are correlated, causing confusion and making data analysis challenging. For instance, iterative models may struggle to converge, and lists of variable importance may become distorted. Another issue arises from the presence of all dummy variables, which can create multicollinearity among them. To address this, one column is dropped intentionally. However, this action, aimed at mitigating multicollinearity, carries the potential risk of elevated p-values, reduced R-squared values, and ultimately rendering the data insignificant.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   With reference to the pair plot, “temp” and “atemp” are the two numerical variables which

are highly correlated 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)**
   Residuals distribution should follow normal distribution and centred around 0(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**
   - September month is the best time to expand because it shows high demand.
   - During good weather, the demand and sales of bike will increase which profits the industry
   - Almost for all the categorical columns, the p-value remains zero which is a good sign.

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation "y = mx + c".

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc.

Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term. Regression is broadly divided into simple linear regression and multiple linear regression.

A). Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
B). Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables. The equation for MLR will be: β1 = coefficient for X1 variable β2 = coefficient for X2 variable β3 = coefficient for X3 variable and so on… β0 is the intercept (constant term).

2. Explain Anscombe's quartet in detail

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasise both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.

- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. It values ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data.

- r = 1 means the data is perfectly linear with a positive slope.
- r = -1 means the data is perfectly linear with a negative slope.
- r = 0 means there is no linear association.

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling? (3 marks)

Feature scaling is a method used to normalise or standardise the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalisation is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
● Standardisation, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalisation, standardisation does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardisation.

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

VIF - the variance inflation factor - The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/ (1-R_1^2). If there is perfect correlation, then VIF = infinity, where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables.
If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1.
So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in "infinity".

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line That's roughly straight. The q-q plot is used to answer the following questions:
● Do two data sets come from populations with a common distribution.
● Do two data sets have common location and scale.