Assignment-Based Subjective Questions

Submitted By: Shwetabh Shekhar

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

- There were 5 categorical variables namely season, workingday, month, weekday, weathersit.

From graphical analysis:

- From the dataset it was observed that bike sales were high during the fall season.
- Bikes were rented more on working days.
- Bikes seem to be rented more in Partly cloudy weather.
- Bike popularity seems to be increasing and is higher in 2019.
- Weekday and month did not give much insight into sales

From Linear regression analysis:

- Bike rent was high on working days and increased by a factor of 0.0246.
- Bike sales were less in the spring season and decreased by a factor of -0.1066.
- Bike sales were higher in September by a factor of 0.0667.
- Bike sales were higher in the year 2019 by a factor of 0.2359.

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

- 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.
- Let's say we have 3 types of values in the Categorical column and we want to create a dummy variable for that column. If one variable is not clean and the other semi_clean, then the last one is obviously unclean. So we do not need a 3rd variable to identify the unfurnished.
- Hence if we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

- Temp, atemp both have the highest correlation coefficient of 0.99.

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

There are four assumptions of linear regression that I verified by performing the residual analysis and plotting graphs.

1. The error terms are normally distributed.
    a. I verified it by plotting a distribution plot of the training data and verifying if they are normally distributed.
2. Check for **Homoscedasticity**
    a. Plotted a scatter plot of the residuals and created a line plot to confirm that residuals have equal or almost equal variance across the regression line.
3. The predicted values have a linear relationship with the actual values.
    a. Plotted the graph and observed **a linear relationship between y_test and y_test_pred.**
4. Linear functional form
    a. Observed that the response variable *y* was linearly related to the explanatory variables *X.*

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

The top three features affecting the sales are as follows:

- temp = 0.3896: higher sales when temperature is high.
- WorkingDay = 0.0238: higher sales on a working day.
- Light rain_Light snow_Thunderstorm = -0.2884: sales decreased when there was rain or thunderstorm.

General Subjective Questions

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

- To understand linear regression in detail we need to understand what regression is.
- Regression is plotting a line that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum.
- Linear regression simply follows this. It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.
- The linear regression model gives a sloped straight line describing the relationship within the variables.
- The straight line formula is y = mx+c. The goal of the linear regression algorithm is to get the best values for m and c to find the best fit line.

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

- It's a group of four datasets that appear to be similar when using typical summary statistics, yet have very different distribution and appear different when graphed.
- Essentially it emphasises on the importance of visualising the data to get a clear picture of what's going on rather than depending on summary statistics.

- It comprises 4 data sets that consist of 11 points. The statistical information for all these four datasets are approximately similar.
- When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities.

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

- Pearson's r is a numerical summary of the strength of the linear association between the variables.
- If the variables tend to go up and down together, the correlation coefficient will be positive.
- If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- The Pearson's correlation coefficient varies between -1 and +1 where:
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association.

Q4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling? (3 marks)

- Scaling is the process used to normalise the range of independent variables or features of data.
- It is performed because Machine learning algorithms like linear regression, logistic regression etc. that use gradient descent as an optimization technique require data to be scaled. This is because the difference in ranges of features will cause different step sizes for each feature.
- Thus to ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.
- Normalisation is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
- Standardisation is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
- In order to determine VIF, we fit a regression model between the independent variables. It is calculated as 〖VIF〗_1=1/(1-R_1^2)
- If all the independent variables are orthogonal to each other, then VIF = 1.0.
- If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
- This happens as it is evident from the formulae that with perfect correlation R2 will be close to 1 and hence makes the VIF value higher.

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

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- It is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. It is formed by
    - Vertical axis: Estimated quantiles from data set 1
    - Horizontal axis: Estimated quantiles from data set 2
- Importance in Linear regression is as follows:
    - The sample sizes do not need to be equal.
    - Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.