Linear Regression Assignment – By Gokul Raj

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?

Ans: From the analysis of the categorical variables, we can infer that certain categories within these variables have a significant effect on the dependent variable (bike bookings - cnt).

For instance, variables like 'season' and 'weather condition' show clear variations in bike bookings across different categories.

For example, bookings tend to be higher in summer and winter seasons compared to spring, and clear weather conditions lead to higher bookings than light rain or cloudy weather.

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

Ans: When creating dummy variables for categorical features, use drop_first=True to avoid problems with the data. This removes one of the categories as reference, so the data is not repeated and can be used more accurately in the model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

Ans: The variable 'atemp' (temperature) has the highest correlation with the target variable (cnt - bike bookings) according to the pair-plot analysis . This means that there is a strong relationship between temperature and bike bookings, and as the temperature rises, the number of bike bookings tends to increase too.

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

Ans: To validate the assumptions of Linear Regression after building the model on the training set, several techniques can be employed:

a. Residual Analysis: Plotting the distribution of residuals to check for normality and zero mean.

b. Scatter Plots: Plotting the predicted values against the actual target values to check linearity.

c. VIF (Variance Inflation Factor) Analysis: To detect multicollinearity among the features.

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

Ans: Temperature (temp): When it's warmer outside, more people like to rent bikes. So, higher temperatures mean more bike rentals.

Year 2019 (yr_2019): In the year 2019, there was an increase in bike rentals compared to other years. This means more people chose to rent bikes in that particular year.

Weather Condition (weathersit_Light Rain): When there is light rain, fewer people rent bikes. So, if it's lightly raining, there are fewer bike rentals.

These three things have the biggest impact on why people decide to rent shared bikes, and knowing about them helps us understand and predict bike rental demands better.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans: Linear Regression is a tool that helps us predict numbers.

Let's say we have some data, like the price of houses, and we want to understand how the size of a house and the number of rooms affect its price. Linear Regression finds a line that best fits the data points and helps us make predictions.

Imagine drawing a straight line through the points on a graph that comes closest to touching each point. This line is our "best-fitted" line. Once we have this line, we can use it to predict the price of a house based on its size and number of rooms.

The algorithm calculates the coefficients (the "weights") of this line using some mathematical tricks, and these coefficients tell us how much each factor (size and number of rooms) influences the house's price.

If the coefficient is positive, it means an increase in that factor will increase the price, and if it's negative, it means an increase in that factor will decrease the price.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's Quartet is like a set of four secret codes that look very different at first glance, but they actually have the same hidden patterns. These four datasets have the same average (mean), the same spread (variance), and even the same straight line when you draw a line through the points.

So, as a data analyst, Anscombe's Quartet reminds us that looking at data visually is important because sometimes things that seem different on the surface may have hidden similarities, and we need to see the big picture to make more accurate and meaningful conclusions.

3. What is Pearson's R?

Ans: Pearson's R is like a measurement tool that tells us how two things are related to each other. It gives us a number between -1 and +1.

If the number is close to +1, it means the two things are positively related, meaning when one goes up, the other tends to go up too.

If the number is close to -1, it means the two things are negatively related, meaning when one goes up, the other tends to go down.

If the number is close to 0, it means there's no clear relationship between the two things.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is like putting different things on the same measuring scale so we can compare them easily.

There are two common scaling techniques:

Normalized Scaling (Min-Max Scaling): This technique scales the numbers to a range between 0 and 1. It's like squeezing all the numbers into a small box between 0 and 1.

Standardized Scaling (Z-score Scaling): This technique makes sure the numbers have zero average (mean) and are spread out in a way that the variance is 1. It's like making sure all the numbers behave in a similar manner around the average.

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

Ans: VIF stands for "Variance Inflation Factor," and it helps us see if there is a problem called "multicollinearity" in our data. Multicollinearity happens when one or more features in our data can be perfectly predicted from a combination of other features.

In a regression model, if we find that some features are perfectly predictable from a combination of other features, it causes the VIF value to become infinite. This is a problem because it makes it impossible for the model to estimate the correct relationship between the features and the target variable.

When we encounter infinite VIF, it's a signal that we need to reevaluate our model and if needed remove some redundant features to avoid this multicollinearity issue and build a better and more accurate model.

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

Ans: A Q-Q plot is like a special graph that helps us see if our data follows a specific pattern, like the normal distribution (a bell-shaped curve). It compares our data to what we would expect if it followed that pattern.

In linear regression, we use the Q-Q plot to check if the "residuals" (the differences between the predicted values and the actual values) follow the normal distribution. If the residuals make a straight line on the Q-Q plot, it means they are normally distributed, and it's a good sign for our regression model. But if they deviate a lot from the straight line, it suggests there might be some problems with our model's assumptions, and we might need to check and adjust it.