Assignment-based Subjective Questions

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

Ans- Categorical variable: weathersit

People tends to take more bikes on Clear and partly cloudy days than cloudy and rainy days.

Categorical variable: Season

People tends to take more bikes in summer and fall season instead of spring and winter season.

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

Ans- Suppose, we have 3 variables A, B & C.

so, if we don't drop any variable, how we give values

A B C

1 0 0

0 1 0

0 0 1

But, if we drop A, we give 0 for both B & C, then A will automatically will be 1.

for ex. B C

0 0

1 0

0 1

So, that's why we drop first variable, and there is a rule to make dummy variables, i.e. n-1

where n = no. of levels

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

Ans- temp variable has the highest correlation with the target variable.

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

Ans- We will predict the model and calculate the residuals and then make a histogram on these residuals. So that we can conclude that error term is normally distributed with mean zero.

We will make a scatter plot on residuals to check for homoscedasticity to know that error terms have constant variance.

We will make a scatter plot on residuals to know error terms are independent of each other.

Q5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Ans- year, weathersit misty and cloudy, weathersit light snow rain and cloudy 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.

Ans- Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

Linear regression equation

In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

In the following image, a linear regression model is described by the regression line y = 153.21 + 900.39x. The model describes the relationship between the dependent variable, Diabetes pregression, and the explanatory variable, Serum triglycerides level. A positive correlation is shown. This example demonstrates a linear regression model with two variables. Although it is not possible to visualize models with more than three variables, practically, a model can have any number of variables.

Linear Regression example

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

Q2. Explain the Anscombe's quartet in detail.

Ans-Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

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. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm.

Q3. What is Pearson's R?

Ans- The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

Understanding the Pearson Coefficient

To find the Pearson coefficient, also referred to as the Pearson correlation coefficient or the Pearson product-moment correlation coefficient, the two variables are placed on a scatter plot. The variables are denoted as X and Y. There must be some linearity for the coefficient to be calculated; a scatter plot not depicting any resemblance to a linear relationship will be useless. The closer the resemblance to a straight line of the scatter plot, the higher the strength of association. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans- 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.

Why?

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.

Normalization/Min-Max Scaling:

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

Standardization 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.

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

Ans- The Variance Inflation Factor (VIF) measures the degree of multicollinearity in a regression model. A VIF value can become infinite in the following scenario:

Perfect Multicollinearity: When one or more predictor variables are perfectly correlated or linearly dependent on each other, it leads to perfect multicollinearity. In this case, one variable can be expressed as an exact linear combination of the others, meaning there's no variation left to estimate its coefficients independently.

Perfect Linear Relationship: If there is a perfect linear relationship between one predictor variable and a combination of other predictor variables, the coefficient of determination (R^2) for that predictor becomes 1 when regressed against the others.

When R^2 is 1, the VIF formula becomes:

$VIF = 1 / (1 - R^2)$

In this scenario, $1 - R^2$ equals 0, causing division by zero in the VIF formula, which results in an infinite VIF.

Infinite VIF values indicate that one or more predictor variables can be perfectly predicted from the others, and there's no meaningful variation left to estimate their coefficients. This situation makes it challenging to interpret the regression model and can lead to unstable parameter estimates.

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

Ans- A Quantile-Quantile (Q-Q) plot, also known as a quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. It's especially useful for comparing a dataset's distribution to a normal distribution. Q-Q plots are commonly used in linear regression and other statistical analyses for the following purposes:

Use and Importance of Q-Q Plot in Linear Regression:

Normality Assumption Check: Linear regression models often assume that the residuals (the differences between observed values and predicted values) are normally distributed. If this assumption is violated, it can lead to biased parameter estimates and incorrect statistical inferences. A Q-Q plot helps assess the normality of residuals.

Residual Diagnostics: By creating a Q-Q plot of the residuals, you can visually inspect how closely the distribution of residuals matches a normal distribution. Deviations from a straight line in the Q-Q plot

can indicate departures from normality.

Detecting Skewness and Outliers: Q-Q plots can reveal the presence of skewness (asymmetry) and outliers in the data. Departures from the theoretical straight line in the Q-Q plot may indicate skewness (curvature) or outliers (discrepancies at the tails).

Model Adequacy: A Q-Q plot is an essential tool for assessing the adequacy of your linear regression model. It helps ensure that the residuals meet the model's assumptions, and the model is valid for making predictions and drawing inferences.

Here's how to interpret a Q-Q plot:

A perfectly straight line represents a dataset that follows a normal distribution. Deviations from this line suggest non-normality.

If the points in the Q-Q plot deviate upward from the line at the ends, it indicates that the data have heavier tails (more extreme values) than a normal distribution.

If the points deviate downward, it suggests that the data have lighter tails (fewer extreme values) than a normal distribution.

S-shaped deviations may indicate skewness in the data.

In summary, Q-Q plots are valuable tools for checking the normality of residuals, detecting skewness and outliers, and assessing the adequacy of a linear regression model. They allow you to identify potential issues that can impact the reliability and validity of your regression analysis, making them a crucial component of the model validation process.