Bike Sharing Assignment by Vyankatesh Kale

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?

Answer :

- More Bikes are rented in the fall season
- More Bikes are rented in the year of 2019
- Bikes are rented more in normal working day than weekend or holiday
- More bikes rented on Saturday
- During the weathersit Clear, Mist+ Cloudy weather bikes rented more

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

Answer : drop_first=True is used to drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.

If we do not use drop_first= True then n dummy variables will be created, and these predictors are themselves correlated which is known as multicollinearity.

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

Answer : The highest correlation with the target variable is 'temp' with correlation of 0.65

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

Answer : Assumptions of linear regression after building the model in the training set we can plot residual distribution after plotting it comes out to be normal distribution with a mean value of the 0. This is the way we can validate the assumptions of linear regression after building the model on training set.

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

Answer :

1. Temp(0.65)
2. Year(0.56)
3. Light snow(-0.23)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer : Linear regression is a type of supervised machine learning algorithm that computes the linear relation between a dependent variable and one or more independent features. When the number of the independent feature is 1 then it is known as Univariate Linear regression and in case of more than one feature it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of dependent variable based on the independent variables. The equation provides a straight line that represents the relationship how much the dependent variable charges for a unit change in the independent variable.

Linear regression is used in many different fields like finance, economy and psychology to understand and predict the behavior of a particular variable. For example in finance linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.

One of the most important supervised learning tasks is regression. In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Linear regression is used to finds a linear equation that best describes the correlation of the explanatory variables with dependent variables. This is achieved by fitting a line to the data using least square the line tries to minimize the sum of the squares of the residuals. The residuals is the distance between the line and the actual value of the explanatory variables. Finding the line of best fit is an iterative process.

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

y = b0 +b1x1 +b2x2 +…

y is the dependent variable

x1,x2 is explanatory variables

b1,b2 is the coefficients

b0 is the intercept that indicates the value of dependent variable

2. Explain the Anscombe's quartet in detail.

Answer : Anscombe's quarter comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line. Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R?

Answer : The Pearson correlation coefficient (***r***) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

(r) Between 0 and 1 – Positive correlation – when one variable changes other also changes in same direction.

(r) is 0 - No correlation - There is no relationship between the variables.

Between 0 and -1 - Negative correlation - when one variable changes other changes in Opposite direction

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

(r) is greater than 0.5 then strength is Strong and direction is Positive.

(r) is between 0.3 and 0.5 then strength is Moderate and direction is Positive.

(r) is between 0 and 0.3 then strength is Weak and direction is Positive.

(r) is between 0 then strength is None and direction is None.

(r) is between 0 and -0.3 then strength is Weak and direction is Negative.

(r) is between -0.3 and -0.5 then strength is Moderate and direction is Negative.

(r) is less than -0.5 then strength is Strong and direction is Negative.

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

Answer : Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Feature scaling is employed for number of purposes:

- Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. You can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.
- Algorithm performance improvement: When the features are scaled, several machine learning methods, including gradient descent-based algorithms, distance-based algorithms (such k-nearest neighbours), and support vector machines, perform better or converge more quickly. The algorithm's performance can be enhanced by scaling the features, which can hasten the convergence of the algorithm to the ideal outcome.
- Preventing numerical instability: Numerical instability can be prevented by avoiding significant scale disparities between features. Examples include distance calculations or matrix operations,

where having features with radically differing scales can result in numerical overflow or underflow problems. Stable computations are ensured and these issues are mitigated by scaling the features.

- Scaling features makes ensuring that each characteristic is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning, producing skewed outcomes. This bias is removed through scaling, which also guarantees that each feature contributes fairly to model predictions.

Normalization	Standardization
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation
When features are on various scales it is functional.	When a variable's mean and standard deviation are both set to 0 it is beneficial.
When the feature distribution is unclear it is helpful	When the feature distribution is considered it is helpful
Values on the scale fall between [0,1] and [-1,1]	Values in a scale are not constrained to a particular range
X_new = (X - X_min)/(X_max - X_min)	X_new = (X - mean)/Std

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

Answer : VIF is a measure of the amount of multicollinearity in regression analysis.

Formula for VIF is :

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R-squared is the coefficient of determination in linear regression. Its value lies between 0 and 1 and 'i' refers to the ith variable.

The greater the value of R-squared greater is the VIF. Hence greater VIF denotes greater correlation . If R-squared value is equal to q then the denominator of the above formula becomes 0 and the overall value become infinite. It denotes perfect correlation

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

Answer : Q-Q plot (Quantile- Quantile) is graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x**.**

A Q-Q plot is a scatter plot created by plotting two sets of quantiles against one another. If the both sets of quantiles are came from the same distribution we should see the points forming a line that is roughly straight. For an example of a normal Q-Q plot when both sets of quartiles truly come from normal distribution