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? (3 marks)

Ans :

season : From the observation we can conclude that maximum bike were sold in season 3 : Fall with median above 5000 followed by season 2 : summer and season 4 : winter. We can conclude that season can be a good predictor of bike sales.

Holiday : Over 97% bike sold was not on Holiday. So we can say that column holiday can not be relevant for a dependant variable

weathersit : We found from the results that around 63% bikes were sold 'weathersit1' with a median of close to 5000 booking (for the period of 2 years). This was followed by 'weathersit2' with 33 % of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

Month : We found variations monthly basis in bikes demand. Month 5,6,7,8 and 9 looks best demanding months. So we can say that month can be a good predictor.

Weekday : We see almost same pattern of bike selling on weekday. This variable can have some or no influence in prediction.

workingday : Almost 69 % of bikes were sold on workingday so it can be a good predictor.

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

Ans :

It facilitates in lowering the more column created throughout dummy variable creation. Hence it reduces the correlations created amongst dummy variables. Get_dummies() function is used to convert categorical variables into dummy variables.

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

Ans : atemp has the highest correlation with target variable with correlation value as 0.65.

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

Ans :

1) The error terms must be normally distributed. Getting normally distributed curve for errors in Residual analysis
2) There should be a linear relationship between dependent (response) variable and independent (predictor) variable(s) which I was able to achieve.

3) The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4) There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation

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

Ans : Temperature , weathersit 3 and year are best features contributing significantly towards explaining the demand of the shared bikes

Temperature (Col_Name -> 'temp') - A coefficient value of '0.549936' indicated that a unit increase in temp variable increases the bike hire numbers by 0.549936 units.

weathersit 3 (Col_Name -> '3') - A coefficient value of '-0.288021' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire

numbers by -0.288021 units.

Year (Col_Name -> 'yr') - A coefficient value of '0.233056' indicated that a unit increase in yr variable increases the bike hire numbers by 0.233056 units.

General Subjective Questions

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

Ans :

Linear Regression is a machine learning algorithm based on supervised learning.

Linear regression is used to predict the outcome of a dependant variable based on independent variables. Here we are getting to specialise in rectilinear regression. Linear regression may be a part of multivariate analysis. Multivariate analysis may be a technique of predictive modelling that helps you to seek out the connection between Input and therefore the target variable.

Linear regression is one among the very basic sorts of machine learning where we train a model to predict the behaviour of your data supported some variables. within the case of rectilinear regression as you'll see the name suggests linear meaning the 2 variables which are on the x-axis and y-axis should be linearly correlated.

For example :

A Real state company wants to predict the rent of newly built flats in a multi store society. Here, Rent of a flat can be predicted based on some historical data having multiple features like Parking, Furnished_status, Airconditioned, area.

Linear regression mathematically- $y = mx + c$

Where,

m - Slope of the line

c – y-intercept of the line

x – independent/predictive variable

y – Dependent/target variable

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

Ans :

Anscombe's Quartet is 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.

The 4 datasets mentioned are:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data

is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by

linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by

linear regression model

Before applying any machine learning method to the dataset, all of the relevant characteristics must be shown so that a good fit model can be created

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

Ans :

Pearson's r is Pearson's Correlation Coefficient, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation.

It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient.

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations

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

Ans :

Scaling is a data pre-processing technique to standardize the independent features present in the data in a fixed range.

The data set contains characteristics with widely disparate magnitudes, units, and ranges. If scaling is not performed, the method simply considers magnitude rather than units, resulting in erroneous modelling. To address this problem, we must scale all of the variables to the same magnitude level.

Normalized scaling: It gathers all of the data between 0 and 1. sklearn.preprocessing MinMaxScaler aids in the implementation of normalisation in Python. Its is also known as MinMax scaling.

MinMax scaling= (x-min(x)) /(max(x)-min(x))

Standardization Scaling: Values are replaced by their Z scores after standardisation. It transforms the data into a conventional normal distribution with a mean of 0 and a standard deviation of 1.

Standardization= (x-mean(x)) / sd(x)

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

Ans :

VIF = Infinite means perfect correlation between 2 variables. VIF is calculated based on correlation co-efficient (R value).

VIF = 1 /1-R2 , For perfect correlation R = 1 which results 1/0 (infinite VIF)

A VIF value of infinity implies that the associated variable may be stated perfectly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans :

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

A quantile is a fraction where certain values fall below that quantile.

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are comparable, the points in the Q–Q plot will roughly correspond to the route y = x. If the distributions are linearly connected, the points in the Q–Q plot will be close to, but not necessarily on, the path y = x. Q–Q plots may also be used to estimate parameters in a location-scale family of distributions graphically.

A Q–Q plot is used to match the shapes of distributions, offering a graphical representation of how features such as location, scale, and skewness differ between the two distributions