General Subjective Questions

1.Explain the linear regression algorithm in detail?

Linear Regression is a supervised machine learning algorithm that executes regression tasks. The prediction values in the model are targeted based on the independent variables. It is mostly used to find out the relationship between variables i.e., independent, and dependent variables and based on the number of independent variables.

Linear regression performs the prediction task, prediction of dependent variable say (y) based on the independent variable say (x). Regression helps to o find out the linear relationship between the input variable(independent variable) and output variable(dependent variable). Hence it is termed as linear regression.

Below given is the hypothesis function of linear regression:

$$\boxed{\boldsymbol{y = mx + c}}$$

Given values:

x= Training input data, independent variable

y= Response variable, dependent variable.

m= Estimated slope (coefficient of x)

c = Estimated Intercept

When training the model, it tries to fit the best line the value of y given x and models gets the best regression fit line by finding m and c values. So, the model used for prediction of y given x.

How to find the best m and c values to get best fit:

To achieve the best fit-regression line, the linear regression model aims t predict y values such that the difference in the error between the predicted value and true value is minimum. So, when we are updating m and c values to achieve the best fit and minimize the error between predicted value and real value. The cost function is described below:

$$\boxed{\boldsymbol{J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2}}$$

Where J is the cost function of linear regression which is the root mean squared error of the predicted value y (predicted y) and real value (y).

Types of Linear Regression:

Two types of linear regression are simple linear regression and multiple linear regression.

If only one(single) independent variable is used to predict the dependent variable, then it is termed as **Simple Linear Regression**.

If more than independent variable is used to predict the de dependent variable, then it is termed as **Multiple Linear Regression**.

2. Explain the Anscombe's quartet in detail ?

Anscombe's quartet can be defined as, if we have a group of four datasets that are nearly identical in terms of descriptive statistics but there are some uniqueness in the dataset that fools the regression model and when they are plotted, we can see that they have different distributions and appear differently when looked at the scatter plots.

In the year 1973 a statistician named Francis Anscombe presented the illustration about the importance of plotting the graphs before analysing and building the model and what are the effects of statistical properties. Four datasets are considered that represented nearly same statistical observations such as mean and variance.

This shows us the importance of visualizing the data before applying the various machine learning algorithms and it also suggests that features must be plotted to see the sample distribution that helps to identifies the anomalies present in the data like outliers, if the data is linearly separable and diversity of the data. When we consider the four datasets that have same statistical properties and when they are plotted on a scatter plot it generates different scatter plots. For example, the datset1 fits the linear regression model pretty good. Dataset2 could not fit linear regression model if the data is non-linear. Dataset3 might have outliers that are not handled by linear regression and dataset4 might also have outliers that are not handled by linear regression.

Hence before implementing any machine learning algorithm the dataset must be properly visualized which will help to make a better fit model.

3. What is Pearson's R?

The Pearson's R or the Pearson's correlation coefficient measures the linear(correlation) relationship between two sets of data , it is a normalized measurement of the covariance, and it always varies between -1 to +1 where:

r=1 indicates the data is perfectly linear with a positive slope(i.e., the two variables are changing in same direction)

r=-1 indicates the data is perfectly linear with a negative slope(i.e., the two variables are changing in different direction)

r=0 indicates there is no relationship(any linear association)

Person coefficient is a good choice if all the below features are available in data

1. Both the variables are numerical
2. The variables are normally distributed
3. The data have no outliers
4. The relationship should be linear

$$\boxed{\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}}$$

Where in the equation:

- Cov(x,y) is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

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

Scaling is the process that comes under data pre-processing when a machine learning algorithm should be applied. In the machine learning a model makes decision according to the dataset applied and calculate the distance between the datapoints to make inferences.

When the data points have features with values closer to each other then the model get trained fast and when the datapoints are far from each other i.e., they have high difference time taken for the model is more and accuracy also be low. So, Scaling is a technique to make the data points come closer to each other. So, that the difference between then will be less. If scaling is not performed, then the model gives higher weights to higher values and lower weights to lower values and which in turn takes more time for training the model.

Normalized Scaling:

In the normalized scaling the values are rescaled in the range 0 to 1. The values will be scaled where the minimum value is 0 and maximum value is 1. It is also known as min-max scaling.

$$\boxed{\boldsymbol{X^{-} = \frac{X - X_{min}}{X_{max} - X_{min}}}}$$

Standardized Scaling:

In Standardize scaling the values are centred around the mean with a unit of standard deviation, the mean of the feature(attribute) becomes zero and we get a distribution that has a unit standard deviation.

$$\boxed{X^{-} = \frac{X - \mu}{\sigma}}$$

Normalized scaling is good to use when we know the distribution does not follow Gaussian distribution. i.e., it is useful in model that does not assume any distribution.

Standardize scaling on other hand is useful when the data flows Gaussian distribution and it is does not have a bounding range, this is useful when there are outliers in data it is not affected by standardization.

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

VIF is an index measure that states how much the variance of an estimated regression coefficient rises due to collinearity. To determine VIF, we need to fit a regression model among the independent variables. In case if we have perfect correlation the VIF is infinity which implies there is a perfect correlation ship between two independent variables. If there is perfect correlation, we get $\boldsymbol{R}^{2}$**=1** which in turn leads to $\frac{\mathbf{1}}{(\mathbf{1}-\boldsymbol{R}^{\mathbf{2})}} = \infty$**.**

To solve the issue of VIF = ∞ one of the variables should be dropped from the dataset that is causing the seamless multicollinearity. If all the independent variables are orthogonal then VIF=1.0 and if VIF=5 this indicated the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity. A general rule is if VIF > 10 then multicollinearity exists. If we have large VIF then following actions can be taken :

1. Review all the independent variables and eliminate that are not adding value to the model
2. Use Principal component analysis that determine optimal set that describes the independent variables
3. Increase the size of sample.

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

Quantile-Quantile(Q-Q) plots is a graphical representation that helps to understand if the dataset plausibly come from any of the distribution such as Uniform, exponential, or Normal distribution. It plots quantiles of the first dataset against the quantiles of other dataset. A 45-degree angle is plotted if the two datasets are coming from same distribution the points will fall on the reference line. It also helps to determine if two datasets are coming from a common distribution. If the two distribution is compared and if they are similar, the points n the Q-Q plot will lie on y=x line, the points in Q-Q plot will lie on the line if the distribution is linear.

Q-Q plots are helpful in linear regression when we receive the training data and test data separately then by using Q-Q plot we can determine if the datasets have same distribution.
Some of the advantages of using Q-Q plots are:

1. The distribution features such as shift in scale and presences of outliers can be found using Q-Q plot.
2. To check if two datasets comes from same distribution
3. Have common scale.
4. Have same distribution and shape

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?**

   a. There is demand in season fall for bike rental compared to rest of the season followed by summer and winter.
   b. Trends from 2018 to 2019 show positive sign which good for business potential after recovery period.
   c. Decline usually happen when weather play is strong role, likewise Light Rain and Snow.
   d. Holidays hasn't contributed much compared to working days.

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

Drop_first= True in significant to use because it helps in reducing the extra column created during dummy variable creation to reduce the correlation that is created among the dummy variables. The main goal is it reduce the number of columns that is not useful, It may not applicable in all the situation when we want to keep first column.

By dropping one of the columns from each categorical feature we can confirm that the remaining columns become linearly independent. When creating a regression model, we can drop one column from each categorical feature. Let us say that we have 3 types of values in categorical column and wish to create dummy variable for that column, then we need 2 columns to represent the dummy variable. So, if we have categorical variable with n-levels, then we need n-1 column that signifies the dummy variable.

The most important thing to be considered here is why we drop one variable in regression? The reason is the importance of that variable can be found by the remaining variables, in order to avoid redundancy, we can drop a column.

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

Temp and a temp has among the numerical variables which has highest correlation with the target variable.

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

For validating the assumptions of linear regression, first thing is to check for the linearity that can be tested using **scatter plot** we can check if linearity is present or not in the data and secondly, for the linear regression analysis needs all features should be multivariate normal.

The multivariate normal assumption can be checked using **histogram** or **Q-Q plot**. When plotting the data points on a Q-Q plot if it forms a straight diagonal line then it is concluded that the assumption is met.

Using the statistical tests like Kolmogorov-Smirnov or Shapiro-Wilk test can be used to check normality error terms from Ordinary Least Square (OLS) . For multicollinearity assumption we can look at the Variance Inflation Factor(VIF).

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

Based on the final model , year , weather and temp are top 3 feature contributing significantly towards the demand of shared bike