Assignment-based Subjective Questions

Ques 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer :

Here many insights can be drawn from the plots:

1. The fall season has the highest demand for rental
2. Demand for next year has
3. Demand is showing continuous growth month on month till September with the highest demand. After September, demand is decreasing.
4. When there is a holiday, demand has
5. Weekday does not give a clear picture about
6. The clear weather situation (weathersit) has the highest
7. During September, bike sharing is During the end and beginning of the year, it is less.

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

Answer:

drop_first=True is important to use, as it helps in reducing the extra columns that get created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Dropping the first columns as (p-1) dummies can explain p categories. In weathersit, the first column was not dropped so as not to lose the info about the severe weather situation.

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

Answer:

Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation (0.63) with the target variable (cnt).

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

Answer:

Residual Analysis:

Errors are normally distributed with a mean of 0. Actual and predicted results follow the same pattern. The error terms are independent of each other.

The R2 value for test predictions:

The R2 value for predictions on test data (0.815) is almost the same as the R2 value of train data(0.818). This is a good R-squared value, hence we can see our model is performing well even on unseen data (test data)

Homoscedacity:

We can observe that the variance of the residuals (error terms) is constant across predictions. i.e., the error term does not vary much as the value of the predictor variable changes.

Plot Test vs Predicted value test:

The prediction for test data is very close to actuals.

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

Answer :

The top 3 features are:

1. yr (positive correlation)
2. temp (positive correlation)
3. weathersit_bad (negative correlation)

General Subjective Questions

Ques 1. Explain the linear regression algorithm in detail.

Answer :

Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis. If there is a single input variable (x), such linear regression is called **simple linear regression**. If there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.

The linear regression model can be represented by the following equation:

y= a0+a1x+ ε

The linear regression model provides a sloped straight line representing the relationship between the variables.

y= Dependent Variable (Target Variable)

x= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best-fit line. The best-fit line should have the least error means the error between predicted values and actual values should be minimized.

The cost function helps to figure out the best possible values for a0 and a1, which provides the best-fit line for the data points. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**.

In Linear Regression, the **Mean Squared Error (MSE)** cost function is used, which is the average of

squared error that occurred between the predicted values and actual values.

Ques 2. Explain the Anscombe's quartet in detail.

Answer :

Anscombe’s Quartet can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

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

Ques 3. What is Pearson’s R?

Answer :

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure of the linear correlation between two variables. It is a number between -1 and 1, where a correlation of 1 indicates a perfect positive correlation, a correlation of -1 indicates a perfect negative correlation and a correlation of 0 indicates no correlation.

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

Examples:

1. **A researcher might use Pearson's R to measure the relationship between the amount of time students spend studying and their grades. A high**

correlation coefficient would suggest that there is a strong relationship between studying and grades.

2. **A marketing manager might use Pearson's R to measure the relationship between advertising spending and sales. A high correlation coefficient would suggest that increasing advertising spending leads to increased sales.**

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

Answer :

Scaling is a step of data Pre-Processing that is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

When we collect data it contains features highly varying in magnitudes, units, and range. If scaling is not done, then the algorithm only takes magnitude into account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.

$$MinMaxScaling = x - min(x)/max(x) - min(x)$$

Standardization Scaling:

- Standardization replaces the values by their Z It brings all of the data into a standard

normal distribution which has mean (μ) zero and standard deviation one (σ).

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

• One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer :

If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables(which shows an infinite VIF as well).

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data sets received separately and then we can confirm using a Q-Q plot that both the data sets are from populations with the same distributions.

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.

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 similar, the points in the Q–Q plot will approximately lie

on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will

approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a

graphical means of estimating parameters in a location-scale family of distributions.