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? (3 marks)

Answer:

- From season point of view, Season 3 has high demand for bikes followed by season2.
- The demand for bikes is high in the Month of Jun to Sep. The Month Jan is the lowest demand month. Thus, Business does not have much reach in the spring season.
- Bike demand is less in holidays in comparison to not being holiday.
- Weekdays have no big impact on business, it has regular demand of 6000 bookings at a constant rate.
- The bike demand is high when the weather is clear and Few clouds (weathersit1), however demand is less in case of Lightsnow and light rainfall (weathersit3).

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

Answer:

- It basically helps reduce the extra columns created during dummy variable creation.
- Let's take an example of this dataset, as we know we have 4 distinct seasons in the data such as Season1, Season2, Season3, Season4. If the season is not in 2,3,4 then it is an obvious assumption that it has to be season 1. Hence to avoid building additional columns and increase size of the dataset, it is recommended to eliminate the first dummy variable.

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

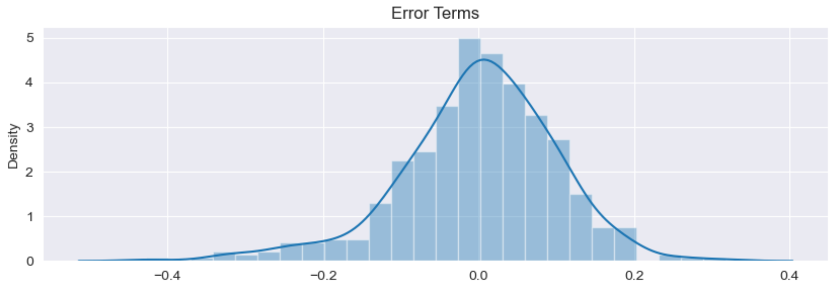
Answer:

- atemp and temp , these 2 have highest correlation with the target variable cnt.

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

Answer:

- By plotting residuals using distplot, we have observed that Linear regression is normally distributed.
- It has a mean of 0.



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

Answer:

Following are the top 3 features contributing significantly towards explaining the demand for shared bikes.

- Temperature (temp) - A coefficient value of '0.5918' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5918 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2595' indicated that, w.r.t Weathersit_2, a unit increase in Weathersit_3 variable decreases the bike hire numbers by 0.2595 units.
- Year (yr) - A coefficient value of '0.2397' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2397 units.

General Subjective Questions

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

Answer:

- It is an algorithm based on supervised learning that represents the relationship between independent (x) and dependent (y) variables to predict the outcome of the upcoming events.
- We can write a linear regression equation as y = .c + mx , where m is slope and c is the intercept of a line.
- There are 2 types of linear regression: simple and multiple linear regression.
- Regression lines are also called the best fit line that fits the given scatter plot in the best way.
- For any best fit line below rules should be followed:
    - There should be linear relationship between X and Y:
    - Error terms should be normally distributed with mean zero(not X, Y):
    - Error terms should be independent of each other:
    - Error terms should have constant variance (homoscedasticity)

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

Answer:

- Anscombe's Quartet is a group of four data sets which are nearly identical in simple descriptive statistics.
- They have very different distributions and appear differently when plotted on scatter plots.
- 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.
- Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson’s R? (3 marks)

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.
- It can be interpreted as follows:
  - Between 0 and 1 - it is considered as Positive correlation & When one variable changes, the other variable changes in the same direction.
  - There is no relationship between the variables if it is 0.
  - Between 0 and -1 - it is considered as Negative correlation & When one variable changes, the other variable changes in the same direction.

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

Answer:

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.
- If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. This could help us to get the accurate coefficients
- Few types of scalings are:
  - sklearn.preprocessing.MinMaxScaler
  - Standardized scaling sklearn.preprocessing.scale

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

Answer:

- Infinite VIF is nothing but perfect correlation between two independent variables
- To avoid such multicollinearity , we should drop one of the variables.

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

Answer:

- Q–Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution.
- If the two data sets come from a common distribution, the points will fall on that reference line.