Rasteh Nili

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)

They have a great impact on explaining variance in renting bikes. Especially weather, holiday season and year. But not all of the categories vary from category to category. One example would be season that summer and winter have significant coefficients with bike rental but in the case of spring it is not true

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

It reduces the extra column created during dummy variable creation. When we generate dummy variables for a categorical column other columns (n-first column) of dummy variables will explain all the variations that exist in the categorical variable. So by removing the first column of dummy variables we prevent correlations created between these variables.

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

temp or atemp depending on which one we decide to keep

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

1. Check linear relationship between x and y: using plot of observed against predicted values (points should be around diagonal line) or residual against predicted values (points should be around horizontal line)
2. Normal distribution of error terms: using distribution plot for residuals or Q-Q plot
3. Independence of error terms from each other: diagnosed using either Durbin-Watson test or residual time series plot or residual autocorrelation plots. The residual autocorrelation falls within the 95% confidence bands around zero.
4. Homoscedasticity: constant variance of error terms diagnosed by plotting residuals versus predicted values or residuals versus time if it is a time series. Residuals that grow larger in either of these cases are violating this assumption.

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

year, light rain and ratio temp from the model used original data and removed VIF.

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm that finds the best fitted line between dependent (y) and independent variables (x). There are two types of linear regression. Simple linear regression that only one independent variable exists and multiple linear regression that model defines the relationship between the dependent variable and multiple independent variables.

Formula for simple linear regression where m is the coefficient or slope and b is intercept.

$$Y = mx + b$$

Formula for multiple linear regression where b_0 is intercept and b_1 to b_n are coefficients or slopes of the independent variables x_1 to x_n.

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

The main aim of linear regressions is finding the optimal intercept and coefficients that minimizes error.

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

Anscombe's quartet are a group of datasets that their mean, standard deviation, linear regression and their simple descriptive statistics are identical but yet have very different distributions and appear very different in graphs. They are used to emphasis graphical exploration of data.

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

Pearson coefficient is a correlation coefficient is the most common way of measuring linear correlation between variables. It is between -1 and 1 that measures both the strength and direction of the relationship between two variables, with 0 being considered as no correlation. It's assumptions are linearity and normality of the data. It is the ratio between the covariance of two variables and the product of their standard deviation as it is shown by the formula below.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

It is only considered the linear correlations and ignores other types of relationship that may exist between two variables.

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

Scaling is transforming the data to make it more suitable for ML models. Different features of data are usually varying in degrees of magnitude, range and units. Scaling can help to improve model performance and reduce the impact of outliers, and ensure that the data is in the same scale and each feature impacts the model equally. It will help gradient descent to converge faster and improve interpretability and comparison of different features.

Normalization is adjusting values of features into a common scale. While standardization makes sure that values of each feature are centered around the mean with a unit standard deviation. Distance based models like PCA are performing better with standardization while Normalization is good for datasets with large outliers and regularly being used for image processing while the loss of outliers is a little disadvantageous in image processing.

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

A large value of VIF indicates a high level of correlation. The higher the correlation the higher the number for VIF. So If VIF is equal to infinity that means the R^2 is equal to 1 that based on VIF formula the denominator of VIF becomes zero. In other words when there is a perfect correlation then VIF will be equal to infinity.

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