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?

Ans) To analyse the effect of categorical variables on dependent variable I made the use of seaborn to plot boxplots and countplots for the respective categorical variable against the dependent variable. By looking at these plots I can say that season plays quite a good role in determining the total bike rented on a day as for the fall season the bike rentals per day are the highest and for the spring season bike rentals per day is lowest. Also in yr 2019 count of bike rented per day are very much high in comparison to yr 2018. Also for month categorical variable total bike rented per day in month 1 is the lowest. For holiday categorical variable also when day is not a holiday then the total count of bike rented per day is somewhat higher in comparison to a holiday. Finally for weathersit variable also the total count of bike rented per day is highest when the weather is clear or less cloudy and the total count of count of bike rented per day is lowest when weather is of light rainfall or light snowfall. For the remaining categorical variables I was not able to draw a significant conclusion.

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

Ans) It is important to use drop_first=True because it helps us in reducing the size of the dataset we are using to train the model by dropping the first column of the dummy variables created and also we should always drop a column because for example suppose we have a variable ’Relationship’ with three levels namely, ’Single’, ’In a relationship’, and ’Married’, we will create a dummy table with 3 rows for all the levels. But we can clearly see that there is no need of having 3 rows for defining three different levels. If we drop a row, say for level ’Single’, we would still be able to explain the three levels. If both the dummy variables namely ’In a relationship’ and ’Married’ are equal to zero, that means that the person is single. If ’In a relationship’ is one and ’Married’ is zero, that means that the person is in a relationship and finally, if ’In a relationship’ is zero and ’Married’ is 1, that means that the person is married.

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

Ans) The variable named atemp has the highest correlation with the target variable.

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

Ans) For validating the assumption that there is no multicollinearity between the independent variables I calculated the VIF value for each variable and only if the VIF value for each variable was less than 5 then only I validated that there is no multicollinearity between the independent variables. Also for validating the assumption that error terms should be normally distributed with mean equal to 0, I plotted a histogram of the error terms and in the histogram it could be clearly seen that the error terms follow normal distribution with mean equal to 0.

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

Ans) Based on the final model the top 3 features contributing significantly are light snow or light rain, spring and windspeed.

General Subjective Questions

Q1) Explain the linear regression algorithm in detail.

Ans) Linear regression is a supervised machine learning method which finds a linear equation that can best describe the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals where residual is the distance between the line and the actual value of the explanatory variable. While drawing the best fit line to the datalinear regression makes several assumptions mentioned below:

1) Relationship between dependent and independent variables is linear
2) The error terms of the model are linearly distributed.
3) Observations are independent of each other.

Q2) Explain the Anscombe’s quartet in detail.

Ans) Anscombe’s quartet consists a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter the plots on a graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Q3) What is Pearson’s R?

Ans) The Pearson’s R is a 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. When the r value lies between 0 and 1 then there is a positive correlation that is when one variable changes, the other variable changes in the same direction and when the r value lies between 0 and -1 then there is a negative correlation that is when one variable changes, the other variable changes in the opposite direction and when the r value is 0 then there is no correlation that is there is no relationship between the variables.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans) Scaling is a method which is used to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying values. Scaling is performed because linear regression algorithm usually assigns a larger weight to higher values and assigns a smaller weight to smaller values, so to avoid this problem and make sure that each feature contributes equally to the learning process we have to scale the features. Standardized scaling basically brings all of the data into a standard normal distribution with mean zero and standard deviation one whereas normalized scaling on the other hand brings all of the data in the range of 0 and 1.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans) When the value of VIF is infinite this means that the R squared score for the predictor variable is 1 that is there is a perfect multicollinearity which means that the predictor variable is an exact linear combination of one or more other predictor variables.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans) Quantile-Quantile (Q-Q) plot, is a graphical tool which helps us in determining 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 came from populations with a common distribution. Q-Q plot helps in linear regression because when we receive training and test data set separately then we can use Q-Q plot to confirm that both the data sets are from populations with same distributions.