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)

- "weekday" – If we consider "cnt" column we do not find any significant pattern with the weekday. However, if the relation is plotted with "registered" users, we observe that bike usage is higher on working days. And with "casual" users it is opposite.
- "season" – The most favorable seasons for biking are summer and fall.
- "weekday" – o If we consider "cnt" column we do not find any significant pattern with the weekday. However, if the relation is plotted with "registered" users, we observe that bike usage is higher on working days. And with "casual" users it opposite.
- "yr" – 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
- "holiday" – Holiday consumption of bikes if compared within "registered" and "casual" users then the observation is "casual" users are using bikes more on holiday.
- "mnth" – The bike rental ratio is higher for June, July, August, September and October months.
- "weathersit" – Most favorable weather condition is the clean/few clouds days. o Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace. o There is no data available for heavy rainy days which means users have not used.

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

Using one-hot encoding the dummy variables are created to cover the range of values of categorical variable. Each dummy variable has 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category. This means if the category variable has 3 categories, there will be 3 dummy variables. The drop_first = True is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

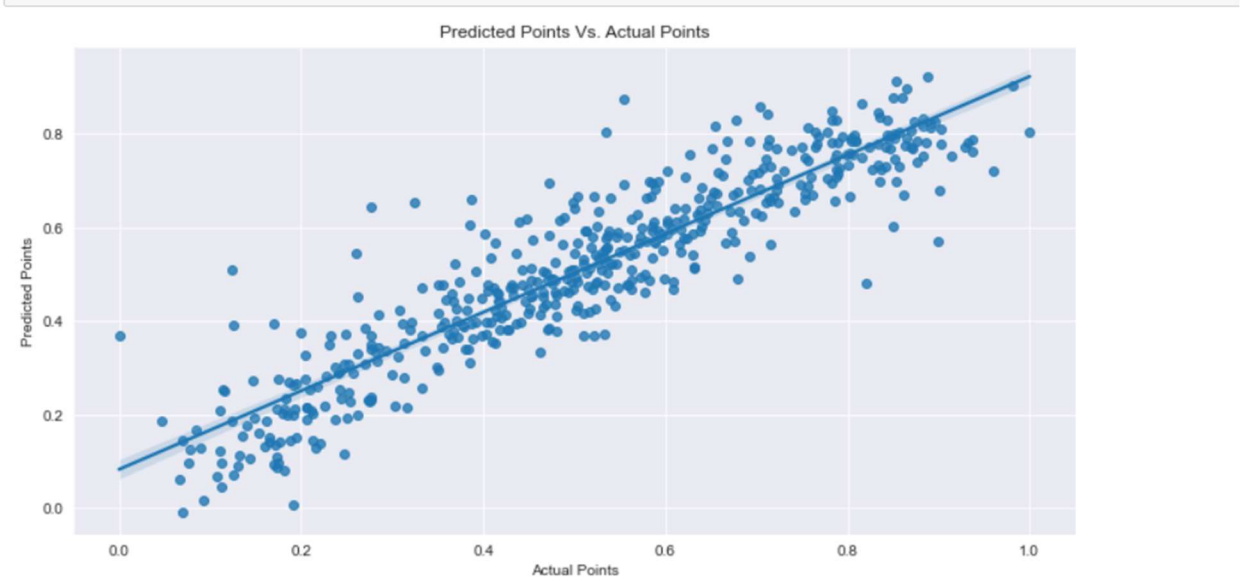
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- "temp" is the variable which has the highest correlation with target variable 0.63

- The casual and registered variables are actually part of the target variable as values of these columns sum up to get the target variable, hence ignoring the correlation of these 2 variables.
- "atemp" is the derived parameter from temp, humidity and windspeed, hence not considering it as it is eliminated in the model preparation.

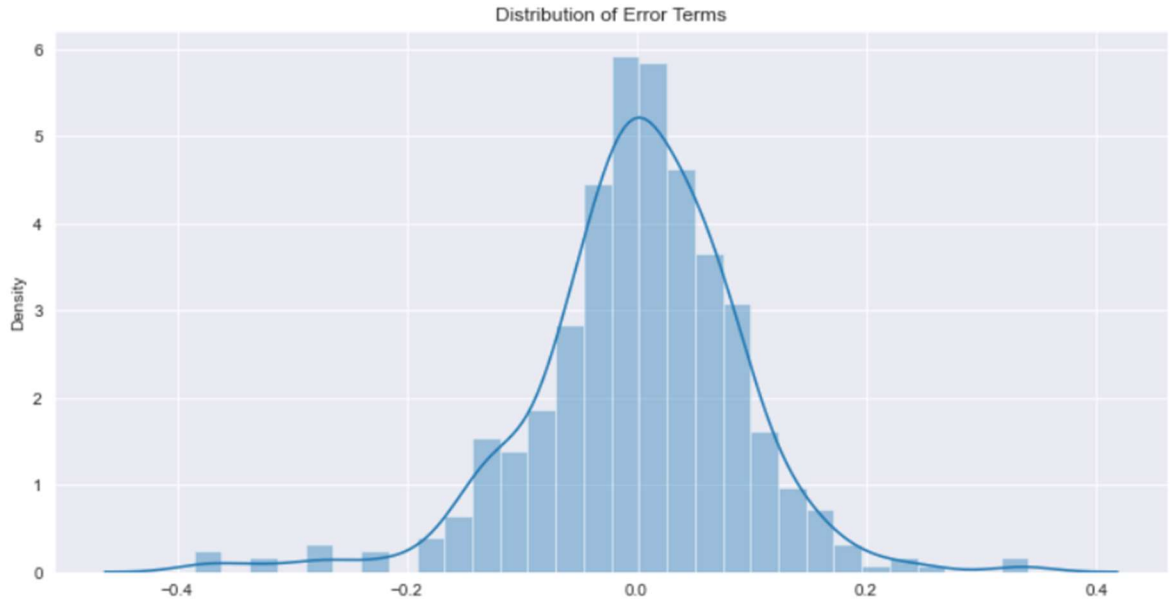
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Linear relationship between independent and dependent variables – A partial residual plot is a graphical technique that attempts to show the relationship between a given independent variable and the response variable given that other independent variables are also in the model.



- Error terms are independent of each other – We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other

- Error terms are normally distributed: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



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

'temp': Temperature is the Most Significant Feature which affects the Business positively, whereas the other Environmental condition such as Raining, Humidity & Wind speed affects the Business negatively. 'Yr': The growth year on year seems organic given the geological attributes. 'season': Winter season is playing the crucial role in the demand of shared bikes

General Subjective Questions

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

Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables. The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

There are 2 types of linear regression algorithms

1. Simple Linear Regression – Single independent variable is used.
   $Y = \beta 0 + \beta 1X$ is the line equation used for SLR
2. Multiple Linear Regression – Multiple independent variables are used.
   $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.
   $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$ o $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$

   Cost functions – The cost functions helps to identify the best possible values for the $\beta 0$, $\beta 1, \beta 2, \ldots, \beta p$ which helps to predict the probability of the target variable.
   The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable.

   There are 2 types of cost function minimization approaches –

   Unconstrained and constrained.
   Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
   The straight-line equation is $Y = \beta 0 + \beta 1X$

   The prediction line equation would be $Ypred = \beta 0 + \beta 1xi$ and the actual Y is as Yi.

$Now\ the\ cost\ function\ will\ be\ J(\beta 1, \beta 0) = \sum(yi - \beta 1xi - \beta 0)$

2 The unconstrained minimization are solved using 2 methods

• Closed form

• Gradient descent

While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used. o $ei = yi - ypred$ is provides the error for each of the data point.

OLS is used to minimize the total e2 which is called as Residual sum of squares.

RSS = = ∑ $(yi - ypred)$ 2 n i=1

Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients

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

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data. Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact

Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.

The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case. Plotting the data is very important and a good practice before analyzing the data. o Outliers should be removed while analyzing the data. o Descriptive statistics do not fully depict the data set in its entirety.

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

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearson's R returns values between -1 and 1. The interpretation of the coefficients are:

-1 coefficient indicates strong inversely proportional relationship.

0 coefficient indicates no relationship.

1 coefficient indicates strong proportional relationship. $r = n(\Sigma x * y) - (\Sigma x) * (\Sigma y) \sqrt{[n\Sigma x\ 2 - (\Sigma x)\ 2] * [n\Sigma y - (\Sigma y)\ 2]}$ Where: N = the number of pairs of scores Σxy = the sum of the products of paired scores Σx = the sum of x scores Σy = the sum of y scores Σx2 = the sum of squared x scores Σy2 = the sum of squared y scores

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

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

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

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and
  1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

- sklearn.preprocessing.scale helps to implement standardization in python.

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

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

The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

VIF= 1/(1-R 2)

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

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 datasets are from common distribution. The theoretical distributions could be of type normal, exponential or

uniform. The Q-Q plots are useful in the linear regression to identify the train dataset and test dataset are from the populations with same distributions. This is another method to check the

normal distribution of the data sets in a straight line with patterns explained below

Interpretations

1. Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.

2. Y values < X values: If y-values quantiles are lower than x-values quantiles.

3. X values < Y values: If x-values quantiles are lower than y-values quantiles.

4. Different distributions – If all the data points are lying away from the straight line.

Advantages

1. Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be identified from the single plot.

2. The plot has a provision to mention the sample size as well.