BIKE SHARING CASE STUDY

Linear Regression Assignment Submission

Student Name: Tarriq Ferrose Khan

Batch : DSC 68

upGrad

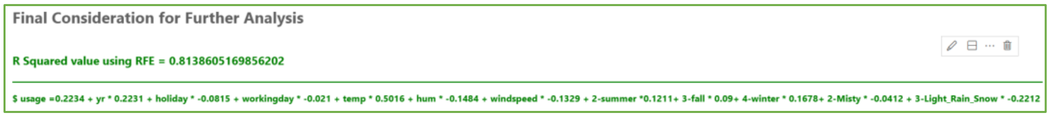
upGrad & IIITB | Data Science Program - May 2024

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:

- R-Squared: 0.813
- Best Fit Equation Screenshot :



Final Consideration for Further Analysis

R Squared value using RFE = 0.8138605169856202

$ usage =0.2234 + yr * 0.2231 + holiday * -0.0815 + workingday * -0.021 + temp * 0.5016 + hum * -0.1484 + windspeed * -0.1329 + 2-summer *0.1211+ 3-fall * 0.09+ 4-winter * 0.1678+ 2-Misty * -0.0412 + 3-Light_Rain_Snow * -0.2212

- Seasons (summer, fall and winter) have positive effect on the number of usages compared to Weather Situations,(Misty, Cloudy etc.,) , humidity and windspeed, which seems to be having negative effect.
- Temperature certainly has positive effect, because in NYC, if the temperature is not too cold or suitable , only then people come out and try using bikes.
- Holiday and Working Day both have negative effect on the expected Usage, may be because NYC Commuters tends to use the tube train on a working day and Holiday there are other venues for entertainment.
- An average of a 22% increase of usage can be expected Y-o-Y , with other variables unchanged.
- If no variables are in effect the expected usage can be = 22%

2. Why is it important to use **drop first=True** during dummy variable creation? (2 marks)

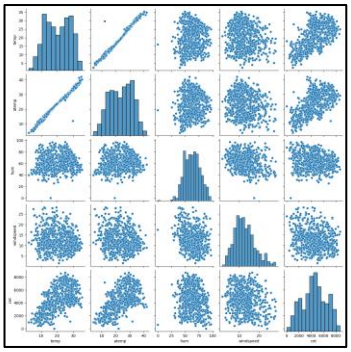
Answer:

- If there are n-variables, to apply scaling , its okay to have n-1 variables created.
- This also prevents multicollinearity among dummy variables.

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

Answer:

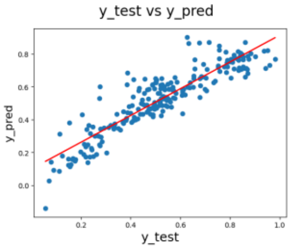
- The variables temp and atemp seems to be having highest correlation with Target variable cnt



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

Answer:

- The test and predicted model are plotted in scatter plot to check if the linear relationship exists.
- Calculated R-Squared which is 0.813 (closer to 1).



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

Answer:

1. Season (Summer, Fall, Winter)
2. Weather Situation (Cloudy, Misty, Light Rain or Snow)
3. Temperature, Humidity.
4. Weekday , Holiday.

General Subjective Questions

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

Answer

- One of the Supervised Machine Learning Technique, which deals with Continuous variables prediction.
- It is used to estimate the Relation ship between one Dependent and one or more independent variables, provided a linear relationship exists between them.
- One Dependent variable, called as Simple Linear Regression, Equation: $y=\beta0+\beta1X$
- Multiple Dependent variables, called as Multi Linear Regression, Equation: $y=\beta0+\beta1X1+\beta2X2+………\beta nXNA$
  - Y is the dependent variable
  - β0 – intercept (where X=0) ,value of Y when X (all X=0, in case of Multi Linear regression)
  - βI –Coefficients - expected change of Y for a one-unit increase in X when all other independent variables) are held constant.
  - X – Dependent variable(s)
- This algorithm intends find the best-fit line, which tells the different between the predicted and actual values should be kept to a minimum.
- **Ordinary Least Squares regression** (**OLS**) method used for estimating coefficients of linear equation, which tells relationship between one or more independent quantitative variables.
- **This is evaluated using** R-squared value which is called Coefficient of Determination, formula = 1- (RSS/TSS)
  - **RSS – Sum of Squared Regression**
  - **TSS – Total Sum of Squares**
- Assumptions of Linear Regression:
  - A linear relationship exists between the Dependent and Independent variable(s)
  - Homoscedasticity: The variance of the error term is the same for all values of the independent variable.
  - Independence: The Independent variables are not correlated, which is checked and eliminated through VIF: Variance Inflation Factor.
  - Normality : The error term is expected to be normally distributed centred at mean=0

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

Answer:

- Emphasizes the importance of EDA and the drawbacks of depending only on summary statistics.
- It emphasizes the importance of plotting data to confirm the validity of the model fit.
- Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quite similar, but when plotted on a graph shows different patterns.

3. What is Pearson’s R? (3 marks)

Answer:

- It is the most common way of measuring the strength and direction of the relationship between two variables.
- It’s the ratio between the covariance of two variables and the product of their standard deviations;
- Ranges between 1 to -1
- A value of 1 indicates the linear equation has all points lie on the regression line and they are perfectly related to each other.

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

Answer:

1. If variables are of different value ranges, it will be difficult to fit into a model. That’s why Scaling is performed to have a range harmony among the variables that it can be easily fit into a model.
2. **Normalized Scaling – aka – Min-Max Scaling** converts all data in the range of 0 to 1.
    a. Formula : x- min(x)/Max(x)- Min(x).
3. Standardized Scaling – brings all data into a standard normal distribution with mean=0 and standard deviation = 1.

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

Answer:

1. It indicates a perfect correlation exists between the Variables, which means R-Squared = 1
2. And Formula for VIF = 1/(1-R,Squared), so if R-Squared is 1 , then the denominator becomes ZERO.
3. So, in that case VIF will be 1/0 = infinity

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

Answer:

1. Quantile-Quantile (Q-Q) plot, is a graphical tool used to check if two data sets come from populations with a common distribution.
2. In case if training and test data set received separately and then this plot can be used to confirm if the data sets are from populations with same distributions.

-x-End of Document-x-