Pragya Shuchi Regression Assignment - II

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?

From the analysis , we see there exists seven categorical variables. Year represented by “yr” and season share a mild positive correlation (close to .5) with the demand of the shared bikes represented by “cnt”.
In the overall model , dummy variables derived from the categorical variables happens to be the most contributing independent variables .
> “light_snow” one of the dummy variable derived from weather situation is having the highest influence in the model (with beta-value = -1.3)
> “light snow” is followed by year which is another categorical variable with a positive linear relationship with the target variable
> This is followed by “winter” – a dummy encoded variable derived from season.

Hence categorical variables (once encoded) have the highest effect on the overall demand of the shared bikes.

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

Dummy variable is one of the ways of encoding categorical variables. When we use pd.get_dummy on a categorical variable with n labels, it yields us n number of columns. However, n-1 columns can give the same level of information to the model .For example – if there are three categories in a column – low ,medium & high and if the two derived dummy variables are Os, we can automatically interpret that’s it’s the high. Thus, we can create sufficient information with n-1 columns.

Hence drop_first = True is useful as it reduces the extra column created during dummy variable creation and reduces the correlations created among dummy variables.

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

Among the numerical variables , atemp has the highest correlation with the target variable. Atemp shows a positive correlation with cnt.

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

Assumptions of linear model validated which have been validated :
1) Error terms are normally distributed : Using a distplot of residual terms, I have tried to understand the distribution of error terms which shows the bell-curve/normal distribution.
2) Error terms have zero mean & constant variance. The distribution of error terms as visualised using distplot is centred around zero-its mean.
3) Error terms don’t follow any pattern : Using a scatterplot of predicted y & error terms we are unable to see any such pattern in the error terms

4) The independent variables are not highly correlated to each other : VIF analysis shows that VIF of all variables is restricted to 3 .
5) There exists a linear relationship between independent and target variable : as represented by the scatterplot analysis

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

Based on the final model & analysing the p-values & beta coefficients, we can say that variables of highest importance are as follows:

1)light_snow (one of the dummy variable derived from weather situation)
2) year ("yr")
3) winter (one of the dummy variable derived from season)

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression algorithm is a supervised machine learning technique where we aim to predict a continuous numerical variable with one or more independent variable. It is used to define the linear relationships between the dependent & independent variables.

For a regression line : (Y) = B0 + B1X1 + E

> B0 is the intercept of the regression lines where as B1,B2,Bn are slopes of the line (widely known as beta coefficients)
> Beta coefficients tell the degree of change in the Y variable by unit change in X1 variable
> The algorithm tries to generate the best-fit lines using OLS(ordinary least square) which minimises the cost function(error) by gradient descent algo or differentiating the cost function .
> R2 is used to measure the accuracy of the algorithm which is also the degree of variance in the target variable explained by the model.

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

Anscombe's quartet is a set of four datasets which have very similar statistical properties but yet they appear when they are graphed.
> These four datasets had 2 variables (x,y) and both the columns in the four datsets displayed equal mean , standard deviation , correlation .
> However , when they were graphed , the four datasets showed different level of linear relationship between the included x & y variable.
> The first data plot showed a linear relation between its x & y var
> The second plot showed a non-linear relation between its x & y var
> The third data plot showed presence of outliers

> The fourth data plot showed that just one high leverage point is enough to generate high correlation between x & y

3. What is Pearson's R?

Pearson's correlation coefficient can be defined as the covariance of the two variables(X,y) divided by the product of their standard deviations.
> The value should lie between 1& -1.
> It is used to measures the linear correlation between two variables(X,y)
> The correlation value can help to understand the strength as well as the direction of the correlation.With negative values saying that the two variables are inversely correalted and positive values indicating a positive correlation

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Feature Scaling is a n important step in data pre-processing where we rescale the variables in the dataset to being all the variables in the same range.
Scaling is important to feed the right information to the model , enhance the interpretability & reduce the computation time.It also helps to sustain the normal distribution amongst the variables. When not rescaled,the variables are in different units & the model might end up generating low Beta coefficients to the high scale variable and we might misunderstand the importance of some variables. Even techniques like gradient descent reduce their computation time when provided scaled data.

There are two popular methods of scaling:
1) Normalisation / MinMax Scaler – All the values in the column are reduced between 0 & 1.It is derived by using x – min(x) divided by max(x) – min (x)
2) Standardisation – All the values in the column are fit in a way that the feature mean becomes 0 & sigma equates to 1. The feature follows a normal distribution.

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

If the value of VIF is infinite , it means that the variance in that variable is being explained by the other independent variables to a great extent and hence it will not add any value to the predictive algorithm.
1) VIF is used to check the multicollinearity existing between the independent variables. It is actually derived from a regression line fitted within the independent variables to understand the degree of variance of one independent variable explained by the other independent variable.

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

The Quantile – quantile plot is a graphical method to determine if two samples of data belong to the same population or not. In a q-q plot , we try to plot a certain quantile of the first data against the same quantile of the second data. It can be used :

1) To test whether two samples have the same distribution shape

2) To test whether the two samples have similar tail
3) To test whether the two samples are from same population

In linear regression, QQ plots can used to test the whether the training and test data set received separately are from populations with same distributions or not.