**Linear Regression – Assignment Q&A’s**

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer:

Categorical variables like

**‘yr’** – helps us in evaluating the demand growth in given year: 2018, 2019.

**Inference**: Demand has risen very well in 2019 when compared to 2018

Chart, bar chart

Description automatically generated

**‘month’** - helps us to check demand variation with each month and then we can correlate to seasonal attribute to get high level inferences.

**Inference**: The demand has risen towards mid-year and tapered towards the end.

Chart, box and whisker chart

Description automatically generated

**‘season’** – helps us to check the demand variance across spring (1), summer (2), fall (3) and winter (4).

**Inference**: The demand is high in summer and fall. Demand is lowest in spring.

Chart, box and whisker chart

Description automatically generated

**‘holiday’** – helps us to check the demand on working day and holidays.

**Inference**: On holidays the demand is low when compared to other days.

**‘weekday’** – helps us to find demand across week in general and predict rush days with respect to weekly demands as well.

***Note:*** With given dataset:

Weekday:

Monday - 6

Tuesday - 0

Wednesday - 1

Thursday - 2

Friday - 3

Saturday - 4

Sunday – 5

**Inference:** The below graph clearly shows the demand is high on day 4 and 5 which are Saturday and Sunday.

Chart, bar chart

Description automatically generated

**‘weathersit’** – helps us to check demand rise and fall during weather conditions like Clear (1), Misty (2), Snow (3) and Heavy Rains (4) etc.

***Note:*** We don’t have data for Heavy Rain (4) days.

***Inference:*** It is evident that on clear days (1) the demand is high and on Snow (3) days the demand is lowest

Chart, box and whisker chart

Description automatically generated

**‘workingday’** – helps us to determine whether it’s a holiday or not, based on this we can check the demand variance.

**Inference:** The demand is high on non-working day when compared to working days.

Chart, bar chart, treemap chart

Description automatically generated

2. **Why is it important to use drop\_first=True during dummy variable creation?**

Answer: drop\_first=True can be used while creating dummy variable as it helps in reducing one column and hence directly reduces the correlations for the created dummy variables.

For example: If there’ a column which denotes male-student (1), female-student (2) and staff (3) and we want to create dummy variables for this column. If we have 0 for male-student and 0 for female-student, then that row of data signifies that person is staff member and not a student. So, we may not need 3rd variable to identify the staff.

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

Answer: ‘temp’ and ‘atemp’ has very high correlation with target variable ‘cnt’.

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

Answer: I have stats model and applied Multiple Linear Regression for bike share dataset.

I have validated by looking at p-value and Variance Inflation Factor (VIF) value for a feature before dropping them.

**High - p, High - VIF** -> should be dropped

**High – Low**:

**High - p, Low - VIF** -> remove these first and re-run the model

**High - VIF, Low - p** -> remove these after the above case.

**Low - p, Low - VIF** -> we can keep this feature

**Conditions applied**: Feature with below values are dropped after applying above conditions:

* VIF > 6.5 dropped (Industry standard practice few consider below 10 is good, but conservative models look at 5 as critical value for banking and healthcare related ones. Here I have chosen average at 6.5 as critical value for VIF)
* p-Value > 0.05 dropped

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

Answer:

1. atemp – feels like temperature has highest positive correlation
2. windspeed – has highest negative correlation
3. yr – year attribute too has correlation as the demand shot up sharply in 2019 compared to 2018.

**General Subjective Q&A**

1. **Explain the linear regression algorithm in detail.**

**Answer:** **Linear regression** is a machine learning model classified as **Supervised Learning**.

Linear Regression performs a task to help us in predict the value of target (dependent) attribute by forming a linear equation with one or many independent attributes. Thus helping us to predict the future values by learning on current data set.

Broadly Linear Regression models are of two types:

1. **Simple Linear Regression (SLR)** – Deals with one predictor (independent) variable and a target variable.

Equation: Y = a + bX

Where X -> independent variable

b -> coefficient or slope of the line

a -> intercept value

Y -> target or dependent variable

Chart, scatter chart

Description automatically generated

1. Multiple Linear Regression (MLR) – Deals with multiple predictor (independent) variable and a target variable.

Equation: Y = a + b1X1 + b2X2 + …. + bnXn

Where X1 - Xn -> independent variable

b -> coefficient or slope of the line

a -> intercept value

Y -> target or dependent variable