Omkar Kurve EML22020036

Kurve.omkar@gmail.com

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?**

- Toatl 6 categorical variables in the dataset.
- Season - Maximum bookings found in season 3
- mnth - the bike bookings were happening in the months 5,6,7,8 & 9
- weathersit- the bike bookings were happening during ‘weathersit1
- holiday - Data is biased
- workingday - the bike bookings were happening in ‘workingday’ with a median of close to 5000 bookings.

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

- Whether to get k-1 dummies out of k categorical levels by removing the first level. It helps in reducing the extra column created during dummy variable creation. Hence it 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?**

- Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

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

- From the lr6 model summary, it is evident that all our coefficients are not equal to zero which means We REJECT the NULL HYPOTHESIS
- **Train R^2 :0.824**
- **Test R^2 :0.820**
- This seems to be a really good model train and test are almost equal so we can validate our assumption.

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

As per our final Model, the top 3 predictor variables that influences the bike booking are:

1. Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
2. Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
3. Year (yr) - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

General Subjective Questions:

1.Explain the linear regression algorithm in detail.

Supervised learning in a supervised learning model, the algorithm learns on a labelled dataset, to generate reasonable predictions for the response to new data. (Forecasting outcome of new data.

In that we have two methods

- Regression
- Classification

Linear Regression tends to establish a relationship between a dependent variable(Y) and one or more independent variable(X) by finding the best fit of the straight line. The equation for the Linear model is Y = mX+c, where m is the slope and c are the intercept.

Also, when dependent variables important more than one so analysing it comes under multiple linear regression.

predict a dependent variable value (y) based on a given independent variable (x). we divide data into two parts:

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

θ1: intercept

θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line.

To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

R-squared is a statistical method that determines the goodness of fit.

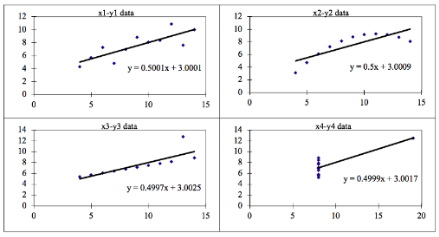
Assumptions of Linear Regression

- Linear regression assumes the linear relationship between the dependent and independent variables.
- Small or no multicollinearity between the features
- Homoscedasticity Assumption
- Normal distribution of error terms.

So, this is the liner regression.

2.Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.



Dataset 1: **this** fits **the linear regression model pretty well.**

Dataset 2: **this** could not fit **linear regression model on the data quite well as the data is non-linear.**

Dataset 3: **shows the** outliers **involved in the dataset which** cannot be handled **by linear regression model**

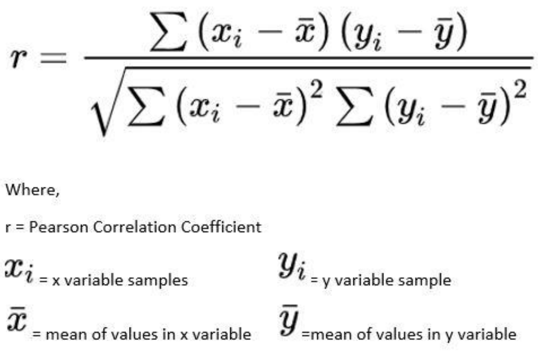
Dataset 4: **shows the** outliers **involved in the dataset which** cannot be handled **by linear regression model**

All the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3.What is Pearson's R?

In statistics, the Pearson correlation coefficient also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

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

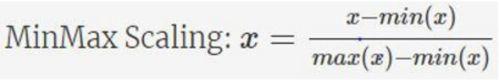
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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:

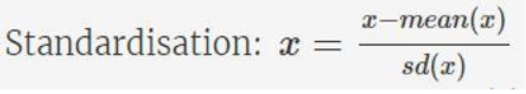
It brings all of the data in the range of 0 and 1.



MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

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).



Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

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

- If all the independent variables are orthogonal to each other, then VIF = 1.0
- If there is perfect correlation, then VIF = infinity.
-

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess 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 come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution