Master of Machine Learning and AI

Machine Learning 1 Assignment

Report

Author:
UpGrad

Student Number:
Phuc Thanh Nguyen

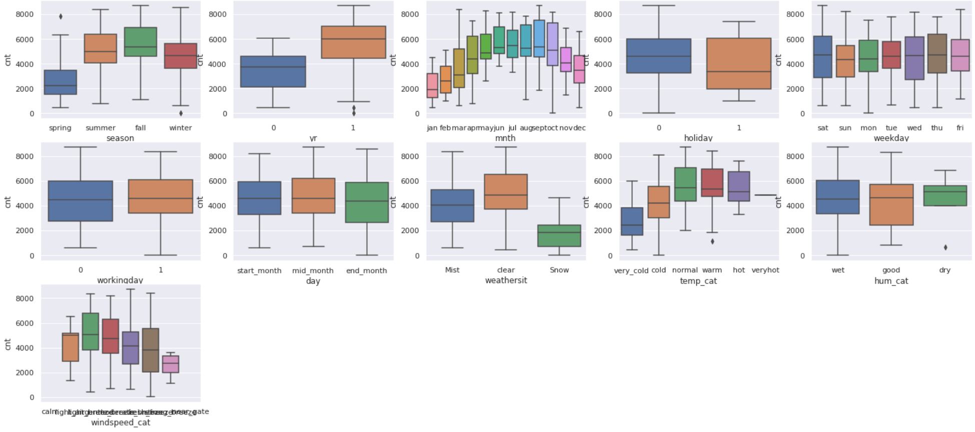
Tuesday 12th April, 2022

Contents

1 Assignment-based Subjective Questions

1.1 Categorical variables affected dependent variable

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Categorical values are considered in dataset, and a few are derived: season, yr , mnth, holiday, weekday, workingday, day, weathersit, temp_cat, hum_cat, windspeed_cat.



Finding from Graph:

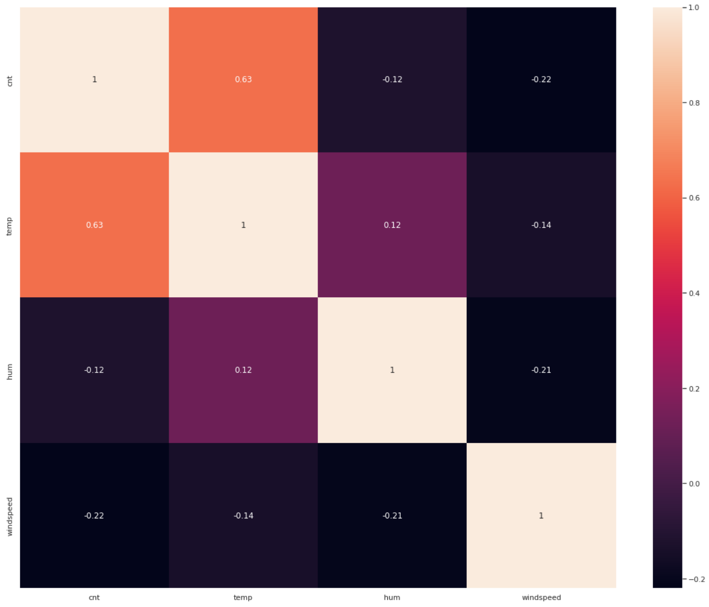
- Demand by season, from high to low: 3 (Fall), 2 (Summer), 4 (Winter) and 1 (Fall), respectively.
- Demand increase in 2019 compared to 2018
- Demand in a year increase from January to September, then fall back.
- Demand is higher when it is holiday
- Weekends have slightly higher demand, but the mean is more or less the same. We can derive that there is not much impact of day of week in the data. And similar is with working day
- End days of month have slightly decrease in number
- Weather impact much in demand: the worse weather is, the lower in demand.
- People would like to have more demand on warm temperate day
- Humidity not affected much for demand
- The demand go highest in light air or light breeze day

1.2 drop_first=True

Why is it important to use drop_first=True during dummy variable creation?
Answer: drop_first=True parameter in the dummy variable creation is important in order to maintain the N-1 Columns,
Example: If we have 12 Columns of the Months, then the final number of columns for dummy variable needs to be N-1 which is achieved by drop_first=True

1.3 Highest correlation

Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer: Looking at the pair-plot , variable with highest correlation is variable "temp" with correlation of 0.63 (as seen in heatmap) with respect to "cnt".



1.4 Assumptions of Linear Regression

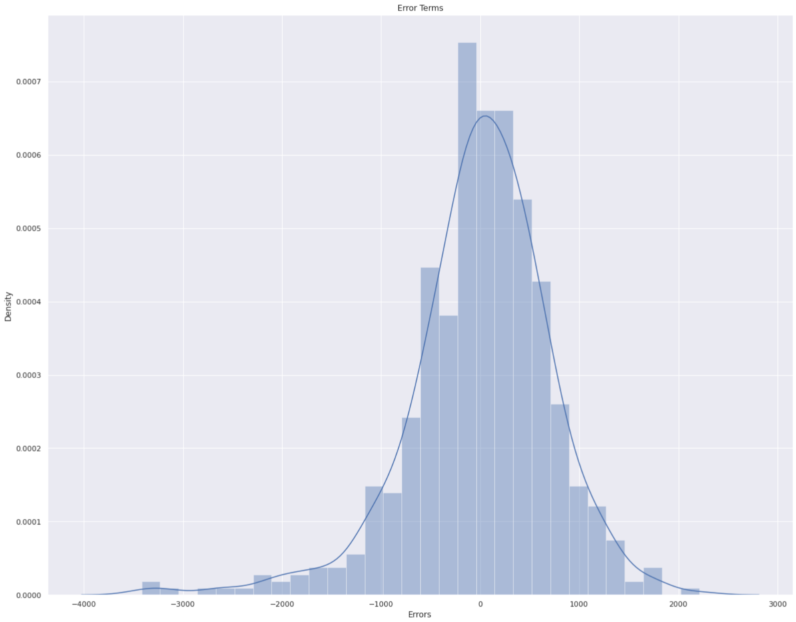
How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: To Validate the assumptions of Linear Regression after Building, it was validated using 4 parameters: Linearity, Muilticolinearity, Homoscedasticity, Normality.

1.4.1 Linearity

Linearity can be tested using scatter plot of independent variables with dependent variables.
By looking at the plots we can see that with the ”cnt variable the independent variables form an accurately linear shape but ”temp” still better than ”windspeed” which seems to hardly have any specific shape. So it shows that a linear regression fitting might not be the best model for it. A linear model might not be able to efficiently explain the data in terms of variability, prediction accuracy etc.

1.4.2 Normality

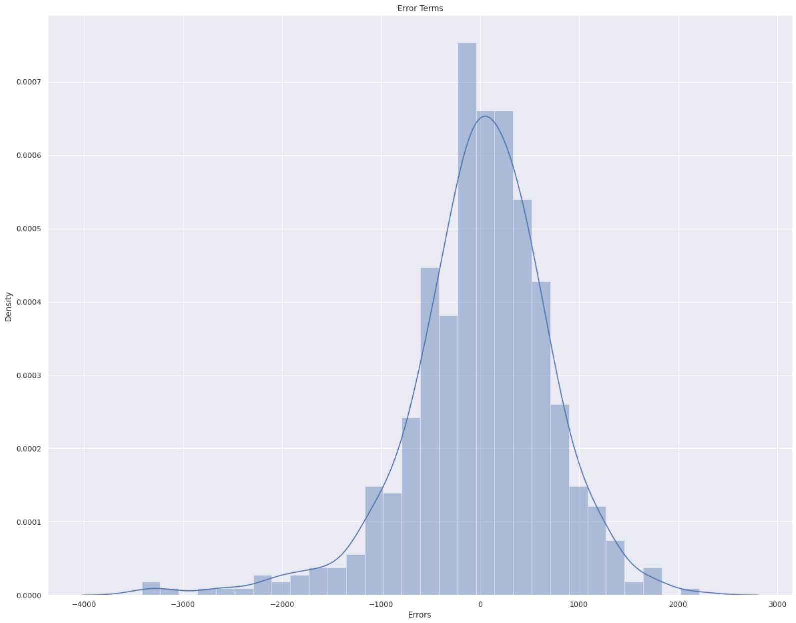
Distribution plot of error term is used to check if they are normally distributed.



Error terms are normally distributed with mean at 0.

1.4.3 Homoscedasticity

Homoscedasticity means that the residuals have equal or almost equal variance across the regression line. By plotting the error terms with predicted terms we can check that there should not be any pattern in the error terms.



This clearly does not look like a constant variance around the zero-line.
Goldfeld Quandt Test: Checking heteroscedasticity : Using Goldfeld Quandt we test for heteroscedasticity.

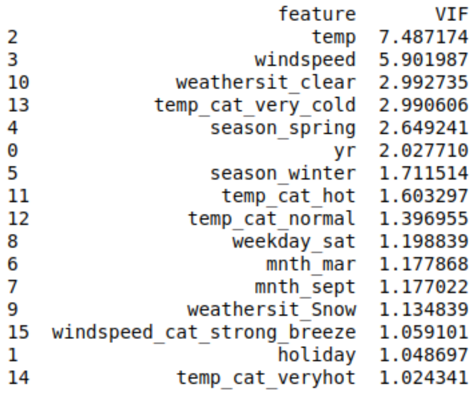
- Null Hypothesis: Error terms are homoscedastic.
- Alternative Hypothesis: Error terms are heteroscedastic.

After running Test:

- F statistic: 1.177149210525878
- p-value: 0.09000220618862494

Since p value is more than 0.05 in Goldfeld Quandt Test, we can't reject it's null hypothesis that error terms are homoscedastic

1.4.4 Muilticolinearity



```
                           feature       VIF
2                             temp  7.487174
3                        windspeed  5.901987
10                weathersit_clear  2.992735
13              temp_cat_very_cold  2.990606
4                    season_spring  2.649241
0                               yr  2.027710
5                    season_winter  1.711514
11                    temp_cat_hot  1.603297
12                 temp_cat_normal  1.396955
8                      weekday_sat  1.198839
6                         mnth_mar  1.177868
7                        mnth_sept  1.177022
9                  weathersit_Snow  1.134839
15  windspeed_cat_strong_breeze  1.059101
1                          holiday  1.048697
14               temp_cat_veryhot  1.024341
```

This data doesn't contain perfect multicollinearity among independent variables, but can be acceptable.

1.5 Top 3 features

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

1. ”temp”: coefficient 3846.07337739
2. ”temp_cat_very_hot”: coefficient -3274.04150401
3. ”yr”: coefficient 2037.54696607

2 General Subjective Questions

2.1 Explain the linear regression algorithm in detail

Linear Regression is a supervised machine learning models that tries to fit a line through a set of data.
Say y is a dependent variable and X is independent variable, which have n attributes. The goal of regression is to find

$$b_0, b_1, b_2, ..., b_n$$

such that, the line denoted by the below equation is best fit line.

$$y = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n$$

The best fit line is the line for which the value of cost function is minimum.
There are many types of cost function, one of them are RMSE.

$$RMSE = \frac{1}{n} \sum_{i=0}^{n} (y_{actual} - y_{predicted})^2$$

Gradient Descent is used to minimize the cost function. It works by iterative finding

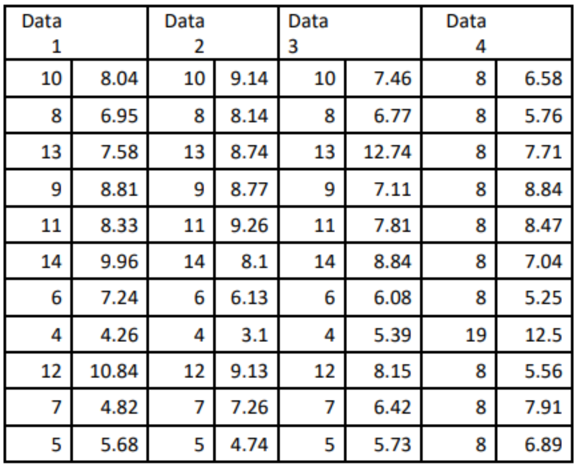
$$b_0, b_1, b_2, ..., b_n$$

values that decreases the cost function till it reaches minimum and no changes occurs.
General steps involved in Building LR model is

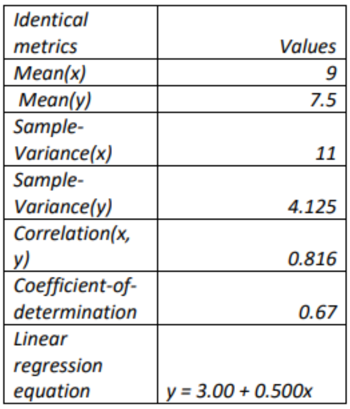
- Data Preparation
- EDA
- Creating dummy variables for categorical data (or one-hot)
- Feature scaling
- Train test split
- Building model with one feature and then adding feature iterative or building model with all features and reducing features by evaluating P-score and VIF
- Residual analysis and Linear Regression assumptions testing
- Model Evaluation and Prediction

2.2 Explain the Anscombe's quartet in detail

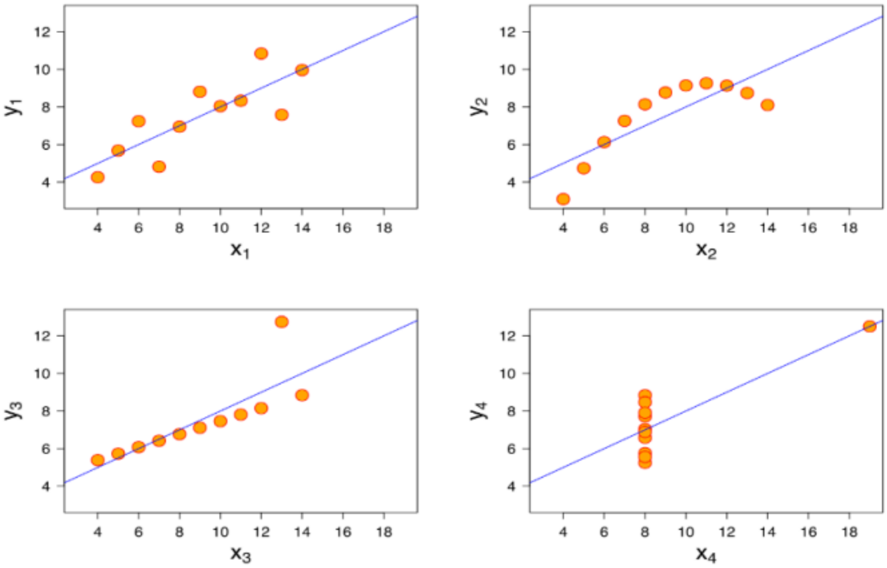
Anscombe's quartet is a combination of four data sets such that they have nearly identical descriptive statistics and yet have very different distributions and look different when plotted.
This dataset was first constructed in 1973 by Francis Anscome. He was a statistician and his findings imposed the importance of visualizing the data as part of analyzing it. It also stated the effect of outliers on the statistical properties.
According to Anscome, he quotes "Numerical Calculations are exact, but graphs are rough"



Data 1		Data 2		Data 3		Data 4	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



Identical metrics	*Values*
Mean(x)	*9*
Mean(y)	*7.5*
Sample-Variance(x)	*11*
Sample-Variance(y)	*4.125*
Correlation(x, y)	*0.816*
Coefficient-of-determination	*0.67*
Linear regression equation	*y = 3.00 + 0.500x*



2.3 What is Pearson's R?

2.3.1 Definition

Pearson's R is the ratio of covariance of two variables and the product of their standard deviations. Hence, it has values between -1 to 1.
It is also widely known as Pearson Product Moment Correlation (PPMC)

2.3.2 Formula

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

2.3.3 Note

PPMC does not take Dependent and Independent Variables into account. This is one of the major drawbacks of PPMC.

2.3.4 Assumptions of PPMC

- Two variables should be measured on a continuous scale (i.e., they are measured at the interval or ratio level). Here, we assume that the variables are not categorical and hence examples like , time series data, Stocks data etc.
- Two continuous variables should be paired,such that each case has two values. Here ”values” can also be defined as ”data points”.
- There should be independence of cases, that is both observations for one case . If observations are not independent, that means the cases are related, and PPMC would not be an appropriate statistical test. There are other sets of Assumptions as well which state the importance of linearity between the variables, Not including outliers, and that there is a need for homoscedasticity.

2.3.5 Values

The Pearson's R lies between -1 to 1 and hence , value of 1 indicates strong correlation, value of -1 indicates strong negative correlation. Whereas, Value of 0 indicates absolute no relation between the variables.

2.4 Scaling

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

2.4.1 Scaling

Scaling is the process of taking a range of numeric values and transforming into a range of different values. It can be thought as changing the units of variables.
2 common scaling method is Min-Max scaling and Standardization.

- Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.
- Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$x_{standardized} = \frac{x - \mu}{\sigma}$$

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

There is no hard and fast rule to determine which is better and depends on application. Standardization is suitable when we know that data follows normal distribution. Normalization helps with outlier treatment. Standardization does not have bounded range. Normalized data will lie between [-1,1].

2.5 VIF is infinite

Infinite VIF means that there is perfect correlation between, the feature and remaining feature. All the variance in the said feature is explained by other feature. VIF is given by below formula: When R-Squared or coefficient of determination for the

$$VIF(X) = \frac{1}{1 - R_{squared(X)}}$$

x feature on remaining features is equal to 1 VIF is infinity.

2.6 Q-Q plot

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

2.6.1 Definition

Q-Q plot stands for Quantile Quantile Plot.
Q-Q plot is a plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

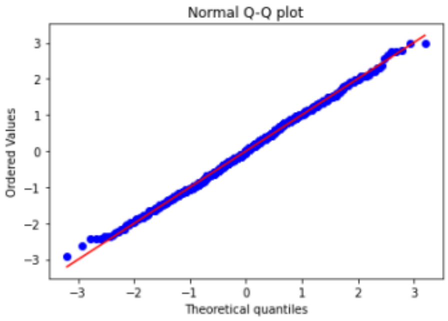
2.6.2 Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

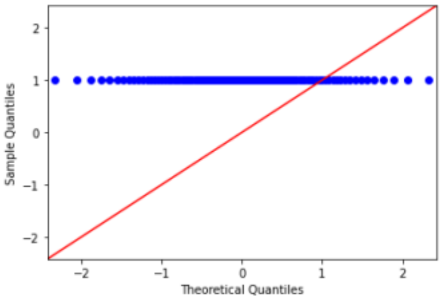
Where, μ is mean and σ is sigma

2.6.3 Usage

Q-Q plots are used to find the type of distribution for a random variable for Gaussian Distribution, Uniform Distribution, Exponential Distribution or Pareto Distribution. The probability plot determines if the data points are to be interpreted as which kind of distribution the data follows. For example, To determine if the data points are Normally distributed, the probability graph would look like this: Similarly, For data which is uniformly distributed, the Q-Q Plot would look something like this:



- To check whether two samples are from the same population.
- To check whether two samples have the same tail
- To check whether two samples have the same distribution shape.
- To check whether two samples have common location behavior.



2.6.4 Q-Q Plot in Linear Regression

Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of properties such as location, scale, and skewness that are similar or different in the two distributions.