upGrad

Answer to Boom Bikes Sales Prediction Modelling and General Linear Regression Subjective Question

- EDA based solution approach

upGrad

Course : Machine Learning

Batch : ML C41

Team : Manish Mandal

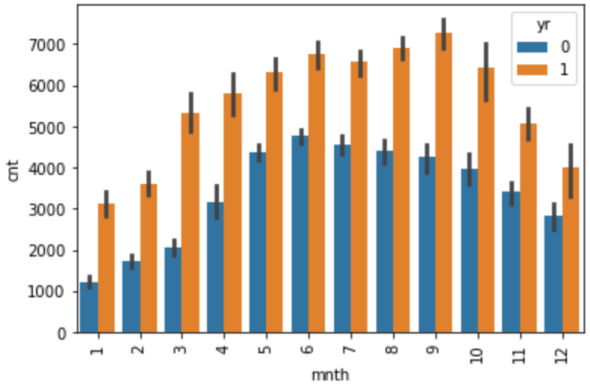
upGrad

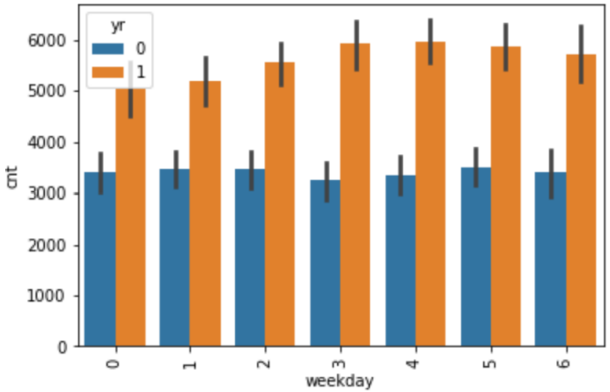
- Answer to BoomBike bike sharing data and prediction model related questions
- Answer to general subjective questions

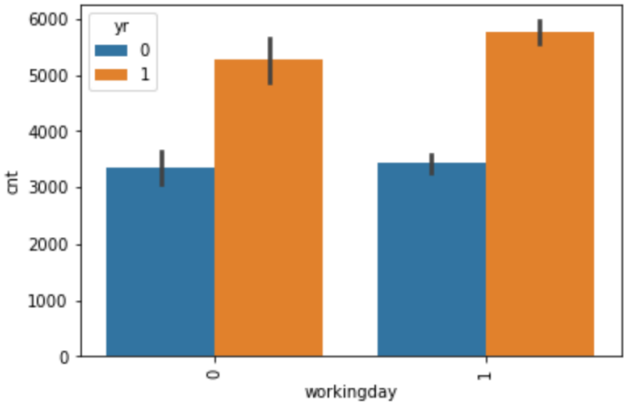
Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer 1:

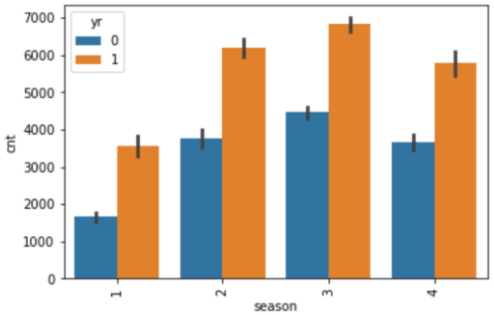
To identify the effect of different key categorical variables, such as “season”, “mnth”, “yr”, “weathersit”, “workingday”, “weekday” on target variable “cnt” different bar plots are visualized for year “2018” and “2019”

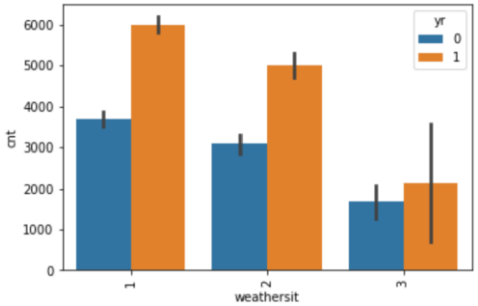






Answer 1: (cntd.)





- The bike consumption has increased from year “2018” to “2019”
- The bike consumption demand increases from Monday to Thursday and dip in demand starts from Friday – as the weekend is approaching with significant drop on Sunday
- The bike consumption is more on working days compared to holiday/ non-working days
- Use of bike is more during spring season this may be due to pleasant spring weather and hence conducive for bike riding
- Number of bike users are more on Clear Sky days relative to other days and significant drop on bad weather days

Q2:	Why is it important to use drop_first=True during dummy variable creation?

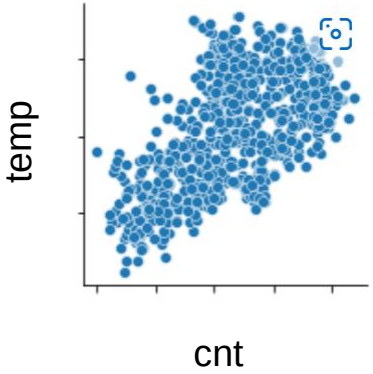
Answer 2:

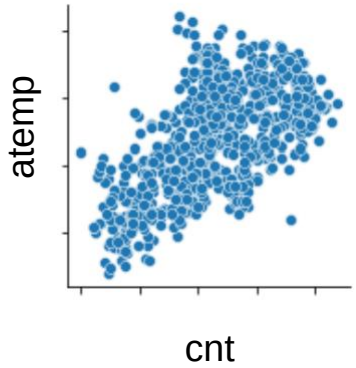
Suppose there are three dummy variables in a given data set.
As x1+x2+x3 = 1 hence anyone can be expressed in relation with other two – like x1 = 1 – (x2+x3).
Thus this will lead to multicollinearity problem.
Hence it is necessary to drop one of them.
The drop_first will drop the first out of n dummy variables hence it is done so. This will help in avoiding dummy variable trap.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer 3:

"temp" or "atemp" (whichever one wants to keep) shows highest linear correlation with the target variable.

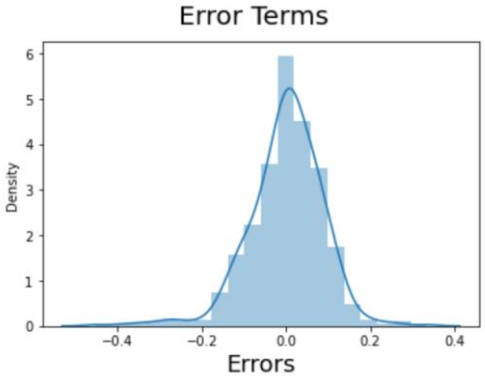




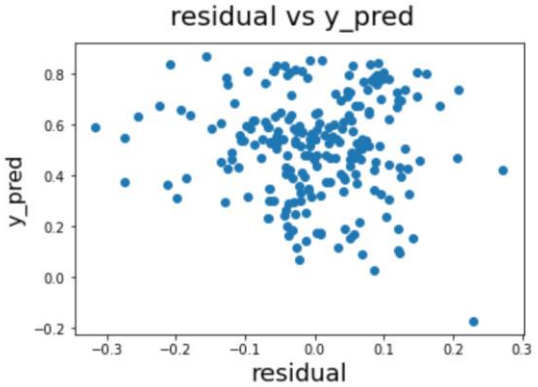
Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer 4:

Following observations validates assumption of Linear Regression



Normal distribution observed in error terms



Homoscedasticity can be observed here as no pattern

Answer 4: contd.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	174.8
Date:	Tue, 02 Aug 2022	Prob (F-statistic):	1.81e-186
Time:	23:18:20	Log-Likelihood:	508.19
No. Observations:	510	AIC:	-984.4
Df Residuals:	494	BIC:	-916.6
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1967	0.033	5.882	0.000	0.131	0.262
yr	0.2357	0.008	29.000	0.000	0.220	0.252
workingday	0.0547	0.011	4.954	0.000	0.033	0.076
atemp	0.4162	0.036	11.473	0.000	0.345	0.488
windspeed	-0.1459	0.025	-5.763	0.000	-0.196	-0.096
spring	-0.0618	0.022	-2.834	0.005	-0.105	-0.019
summer	0.0523	0.016	3.208	0.001	0.020	0.084
winter	0.1028	0.018	5.617	0.000	0.067	0.139
aug	0.0412	0.018	2.327	0.020	0.006	0.076
dec	-0.0518	0.018	-2.949	0.003	-0.086	-0.017
jan	-0.0569	0.018	-3.117	0.002	-0.093	-0.021
nov	-0.0483	0.019	-2.568	0.011	-0.085	-0.011
sep	0.0889	0.018	5.078	0.000	0.055	0.123
cloudy_misty	-0.0842	0.009	-9.676	0.000	-0.101	-0.067
light_snow_and_snow	-0.2943	0.025	-11.981	0.000	-0.343	-0.246
sat	0.0658	0.014	4.624	0.000	0.038	0.094

Omnibus:	83.674	Durbin-Watson:	2.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	247.261
Skew:	-0.777	Prob(JB):	2.03e-54
Kurtosis:	6.037	Cond. No.	21.5

Model with significant variables p < 0.05 for all variables

	Features	VIF
2	atemp	6.38
1	workingday	4.78
3	windspeed	4.63
4	spring	3.22
6	winter	3.17
5	summer	2.54
0	yr	2.07
14	sat	1.87
10	nov	1.78
7	aug	1.72
9	jan	1.67
12	cloudy_misty	1.59
8	dec	1.48
11	sep	1.39
13	light_snow_and_snow	1.09

VIF is low and atemp cannot be ignored – hence assumed absence of multicolinearity

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer 5:

- atemp (+ve)
- Yr (+ve)
- Light_snow_and_snow (-ve)

upGrad

Q1: Explain the linear regression algorithm in detail.

Answer 1:

- atemp (+ve)
- Yr (+ve)
- Light_snow_and_snow (-ve)

Q1: Explain the linear regression algorithm in detail.

Answer 1:

In machine learning the mathematical equation, aka model, is derived from the given sample data set. This derived equation is used to estimate/ determine outcome from the actual population/ test data set. There are different statistical method used to derive the relationship/ equation that can be generally applied on the similar type of data set. One such technique is Regression Analysis to understand the relationship between outcome (target) variable and independent (predictor) variables.

There are different types of regression methods applied to derive the relationship, e.g., Linear Regression, Logistic Regression, Polynomial Regression etc. These different techniques are applied either to predict a value or classify an outcome.

Linear Regression Analysis is one type of supervised machine learning technique to predict values of continuous target variable. As the name suggests the equation for Linear Regression is $y = mx + c$, where y is the target variable and there is only one independent variable x. This is a equation of line where m is the slope and c is the intercept.

If outcome 'y' can be described using only one independent variable 'x'. then Simple Linear Regression method is adopted to determine the best fit value of c (the constant/ intercept) and m (coefficient/ slope). But for almost all cases the outcome is dependent on many factors and to derive the actual best fit equation/ relationship between outcome and corresponding factors the equation takes form like: $y = c + m_1x + m_2x + m_3x + \ldots + m_nx$, which represents a hyperplane. So Multiple Linear Regression technique is applied to derive best fit equation coefficients (m values and c value).

upGrad

Answer 1: contd.

A typical example is that the sales estimate of apparel could be dependent on many factors such as Market Condition, Season, Festival, Income etc.

Following are the key assumptions of Linear Regression:

There are four assumptions associated with a linear regression model:

1. **Linearity**: Relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: Variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other and there is no mulitcolinearity
4. **Normality**: For any fixed value of X, Y is normally distributed.

Q2: Explain the Anscombe's quartet in detail.

Answer 2:

Anscombe's Quartet comprises of group of four datasets having almost same identical simple descriptive statistics, but when plotted they exhibit different relationships between "x" and "y". The descriptive statistics, mean, standard deviation and correlation between "x" and "y" are identical in nature. So to avoid such fool data set situation, it is important to find correct relationship between variables through visualization prior to analysing and model building.

For example following data set is Anscombe's Quartet:

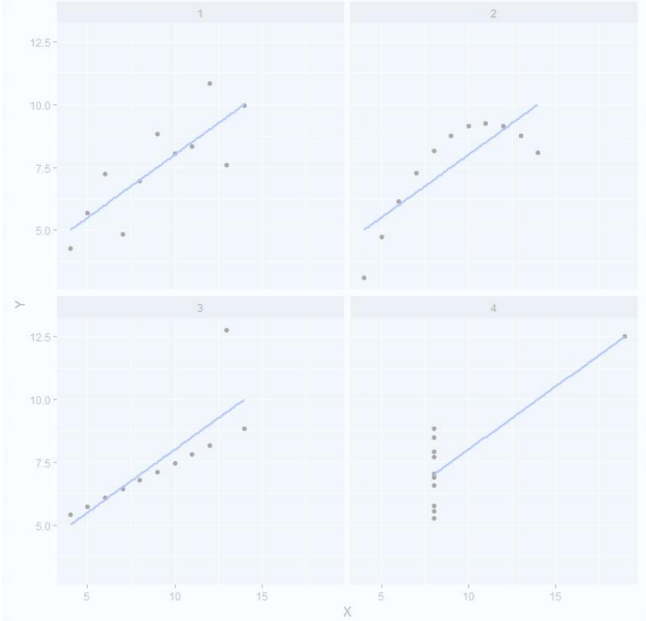
x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

	mean(X)	Std. Dev (X)	mean(Y)	Std. Dev (Y)	Corr(X,Y)
Set 1	9	3.32	7.5	2.03	0.816
Set 2	9	3.32	7.5	2.03	0.816
Set 3	9	3.32	7.5	2.03	0.816
Set 4	9	3.32	7.5	2.03	0.817

upGrad

Answer 2: contd.

Below scatter plot between different sets clearly shows although



Acknowledgment:

Data and Plots are taken from https://www.geeksforgeeks.org/anscombes-quartet/

Q3: What is Pearson's R?

Answer 3:

Pearson's R is aka Pearson's Correlation Coefficient is a measure of bivariate correlation between two variables. It indicates linear correlation between two variables and correlation value can range between -1.0 to +1.0.
The two variables can be of any combination, i.e., it can be between two independent variables or between independent and dependent variables.
The formula is defined as below:

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

Where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer 4:

Most of the times the data received for building models contain variables with varying magnitude of values. This might lead to vary weird coefficient and interpretation of such coefficient is difficult. So it is important to bring all the variables under same scale so that resultant model converges faster and the interpretation is easier.

There are two types of Scaling a. Standardization and b. Normalization

Standardization: In this the variables are scaled in a way so that their mean is zero and standard deviation is 1.

$$z = \frac{x - \mu}{\sigma}$$

Normalization: In this method the scaling is done in a way so that the values of the variables lie between 0 to 1. MinMax scaler is used to do this.

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

Scaling affects coefficients but does not affect any of the other parameters like t-statistic, F statistic, p-values, R-square, etc

upGrad

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer 5:

Variable Inflation Factor (VIF) helps in measuring collinearity between different variables. This is useful to check multicollinearity among different independent variables. If the VIF value is high for an independent variable then it indicates that the variable can be represented by combination of other independent variables in the provided data set. The formula of VIF is 1/(1-R-square).
So if the value of R-square is one then the VIF value can be infinite indicating that there is perfect linear relationship exists for the variable in concern that can be expressed by linear relationship with others.

upGrad

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer 6:

Q-Q plot aka Quantile-Quantile plot, are plots of two quantiles against each other.
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.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties :
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior
A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

upGrad

#LifeKoKaroLift

Thank You!