Bike Rental Prediction

LINEAR REGRESSION

By:

Tejas Kumar D A

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? (3 marks)

- During **Summer ,Winter** season the bike rental demand tend to increase as per the prediction
- Whenever there is **holiday** the demand for bike rental demand reduces
- Whenever there **is light rain** weather situation the bike rental demand reduces
- Whenever there is **windspeed** is more the bike rental demand reduces as it is feasible to ride during this time

	Features	VIF
2	temp	3.63
3	windspeed	2.97
0	yr	2.00
4	summer	1.55
5	winter	1.35
7	Sep	1.20
6	Light Rain	1.06
1	holiday	1.03

Dep. Variable:	cnt	R-squared:	0.803
Model:	OLS	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	256.0
Date:	Wed, 14 Feb 2024	Prob (F-statistic):	1.42e-171
Time:	19:43:16	Log-Likelihood:	453.37
No. Observations:	510	AIC:	-888.7
Df Residuals:	501	BIC:	-850.6
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0872	0.017	5.001	0.000	0.053	0.121
yr	0.2337	0.009	26.078	0.000	0.216	0.251
holiday	-0.0871	0.028	-3.070	0.002	-0.143	-0.031
temp	0.5687	0.021	26.559	0.000	0.527	0.611
windspeed	-0.1453	0.027	-5.325	0.000	-0.199	-0.092
summer	0.0802	0.011	7.140	0.000	0.058	0.102
winter	0.1275	0.011	11.318	0.000	0.105	0.150
Light Rain	-0.2541	0.027	-9.514	0.000	-0.307	-0.202
Sep	0.0891	0.017	5.198	0.000	0.055	0.123

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

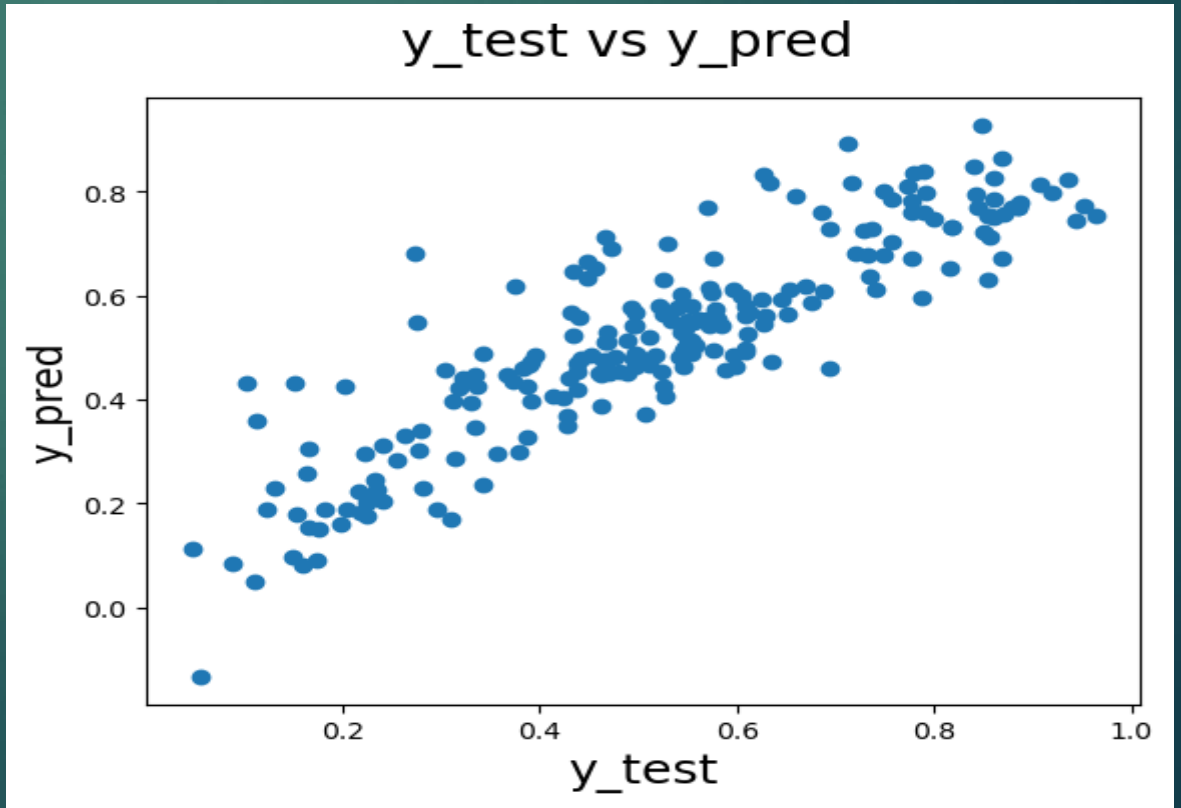
- If there are n dummies n-1 no of dummies represent all dummies ,the dropped dummy can be identified based on the n-1 dummies
- for example, if there are 4 seasons - Spring, Summer ,Winter,fall .Spring can be identified based on last columns Summer,Winter, fall and left out is Spring
- It also reduces the extra column and also reduces correlation among variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- **Temp** variable has the highest correlation with count of Rental bikes

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Using the residual analysis the error term which is the difference of y train and y predicted should follow the normal distribution ie mean of error terms is zero
- Check for Homoscedasticity ie have constant variance of error terms by plotting scatter plot(Y Pred vs Y test)
- Error terms are independent of each other
- No Multi collinearity



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

- Summer and Winter Season
- Temperature
- Year and Holiday

We can see that the equation of our best fitted line is:

cnt = 0.5687×temp+0.2337×yr+0.1275×Winter+0.0891xSep+0.0802×Summer-0.0871×*h*oliday-0.0143×windspeed-0.2541×Light Rain

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method. The output is a continuous variable .

Linear regression is of the 2 types:

a)Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

Formula for the Simple Linear Regression:

$Y=\beta 0+\beta 1X1 +\epsilon$

b)Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression:

$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by the following two methods:

· Differentiation

· Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression

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

Anscombe’s quartet contains 4 graphs which is a graphical representation of datasets that have identical descriptive statistics such same mean,Standarddeviation,Variance etc

Basically, it explains the below:

a) 1.Linearity

b) 2.pearson co efficient.

c) 3.Effect of outliers

d) 4.Correlation coefficient

Thus conveying the visualisation of data's importance with graphical presentation

3. What is Pearson’s R? (3 marks)

Pearson’s R is a correlation coefficient that measures linear correlation between two variables.

It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

a) Pearson’s R is '0' No correlation between variables

b) Pearson’s R between '0' and '1' - Positive Correlation - when one variables change the other changes in same direction.

c) Pearson’s R between '0' and '-1' -Negative Correlation - when one variables change the other changes in opposite direction.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a preprocessing technique applied for normalising the independent varibales which are on a different scale.

Inorder for algorithm to consider magnitude and units of independent variables scaling is perfomed to bring all

Normalised Scaling:

it brings all the data between 0 and 1

sklearn.preprocessing.MinMaxScaler - fit_Transform

Fit- Learns minimum and Maximum

Transform- applies Normalisation using the formula - (x-xmin/xmax-xmin)

The outliers are lost due to this

Standardized 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)

formula x=(x-x(mean))/sd(x)

The outliers are retained

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

(3 marks)

VIF Infinite represents perfect correlation between variables

VIF=1/(1-R2)

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite.Hence indication of perfect correlation

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

Q-Q (quantile-quantile) plots play a vital role in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions that we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight line y = x. A Q-Q plot tells us whether a data set is normally distributed.

Advantages:

1.The sample sizes do not need to be equal.

2.Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

3.The q-q plot can provide more insight into the nature of the difference than analytical methods.

END