Oct 10, 2022

Walmart Sales

What Happens When Wal-Mart Comes to Town?

You have been hired as a consultant for a major local grocery store. Store management is worried since Wal-Mart has entered the market by opening a "Wal-Mart Super-centre" only 3 miles away from the local store. Management is interested in analysing the impact on store sales of the Walmart entry and whether or not a new strategy is required.

For analysis, management has given you access to one hundred weeks of sales data for the local store covering the period both pre- and post-entry of Wal-Mart. Look at the data in Walmart Data.csv

It has the following variables:

WEEK	Week number
Sales	Weekly Sales
Promotion Index	Index of weekly promotion activity –higher promotion index indicates more products on promotion in the store
Walmart	Walmart dummy = 1 in the weeks after the Walmart opens, and 0 in the weeks before the Walmart opens
Feature Advertising Index	Index of feature advertising activity – higher feature advertising index indicates more feature advertising
Holiday	Holiday Dummy = 1 during major holiday weeks, and 0 for non-holiday weeks

Q1. What does the company should do about the possibility of the local store using promotional activity to fight Wal-Mart? What strategy would you recommend to the local store? Develop the appropriate regression model and interpret.

Objective - To identify the impact of all the independent variables(Promotion index, Feature Advertising index) on the dependent variable (Sales)

Justification - Since the dependent variable and all the independent variables are quantitative in nature, we will use regression analysis.

Data Analysis:

Step 1 : Model Testing

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	7.34E+11	3.67E+11	10.82865655	5.69E-05
Residual	97	3.29E+12	33887722216		
Total	99	4.02E+12			

Here p < alpha so we reject the null hypothesis and accept H1. Therefore we can say the model is significant.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.427223401
R Square	0.182519834
Adjusted R Square	0.165664573
Standard Error	184086.1815
Observations	100

Step 2:

Interpretation of R2 (Coefficient of determination): Here R2 value is 0.18 that is below 0.50 that means the model is not very good. The value explains only 18% variation in the range of dependent variable with respect to changes in the independent variable

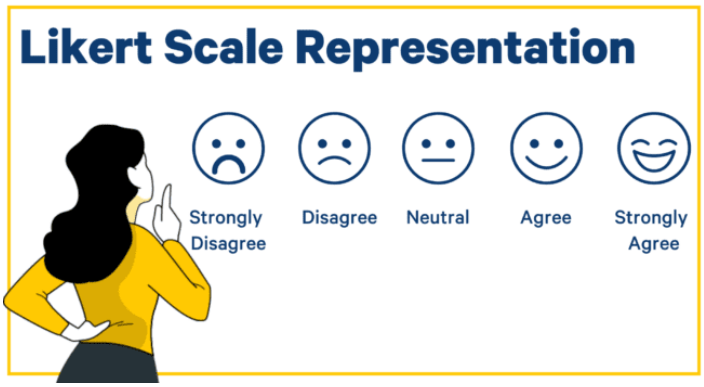
Note *- R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.*

H_0 **(Null Hypothesis) -** The model is not statistically significant.

H_1 **(Alternative Hypothesis) -** The model is statistically significant.

Nov 7, 2022

Likert Scale→ a scale used to represent people's attitudes to a topic. It is quantitative in nature



Case - Experiential Retailing: Influence on young Indian Consumer's response.

Multivariate Regression - atmosphere

Objective - To identify the impact of all independent variables that are sound, light layout, music, fragrance, etc. on the customer retail experience.

Justification - All the variables are quantitative(numerical) in nature. Hence, to check the above mentioned objective, we will use a multivariate regression model.

Data Analysis:

Step 1 : Hypothesis for multivariate linear regression model

H_0 (Null Hypothesis) - The overall model is not statistically significant.

H_1 (Alternative Hypothesis) - The overall model is statistically significant.

Multiple Linear Regression

Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

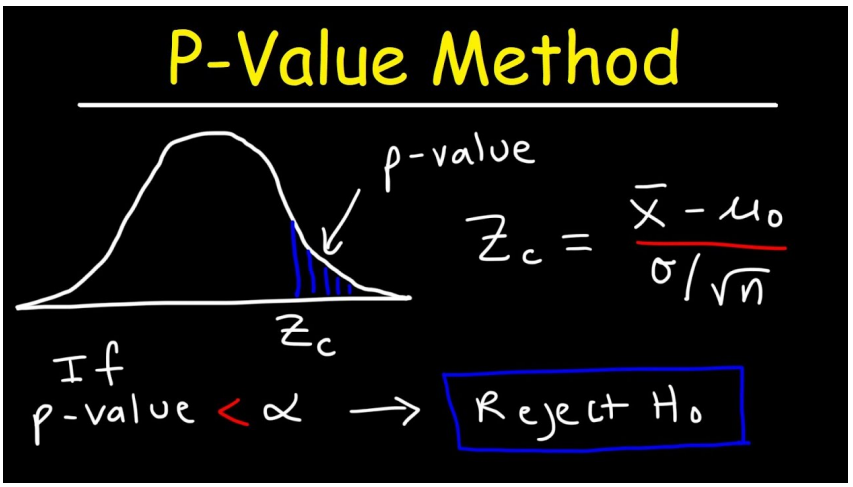
beta_0 is multiplied by 1 for each observation

Can be written in mat $+ \boldsymbol{\epsilon}$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

observations model matrix

model parameter vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$



Coefficients:						
	Estimate	Std.Error	t	p-value	Significant/Insignificant (alpha = 0.05)	
(Intercept)	0.818982	0.707732	1.157	0.250096		
shoping.when.bored	0.009209	0.089835	0.103	0.918567	p>α => Accept Null Hypothesis	Not Significant
waste.of.time	-0.10244	0.110548	-0.927	0.356477	p>α => Accept Null Hypothesis	Not Significant
fragrance	0.084468	0.110294	0.766	0.445669	p>α => Accept Null Hypothesis	Not Significant

wall.colour	0.175777	0.099336	1.77	0.080015	p>α => Accept Null Hypothesis	Significant at 10 percent
emp.knowledge	-0.36703	0.1491	-2.462	0.015635	p<α => Reject Null Hypothesis	Significant
layout.flooring	-0.07854	0.105669	-0.743	0.459183	p>α => Accept Null Hypothesis	Not Significant
recommend	0.044022	0.124292	0.354	0.723986	p>α => Accept Null Hypothesis	Not Significant
emp.concerned	0.176481	0.122625	1.439	0.153383	p>α => Accept Null Hypothesis	Not Significant
layout.spacious	-0.15116	0.099258	-1.523	0.131102	p>α => Accept Null Hypothesis	Not Significant
emp.trustworthy	0.150259	0.116543	1.289	0.200424	p>α => Accept Null Hypothesis	Not Significant
layout.design.display	0.030934	0.122993	0.252	0.801963	p>α => Accept Null Hypothesis	Not Significant
entertain	-0.0762	0.162124	-0.47	0.639416	p>α => Accept Null Hypothesis	Not Significant
enthusiam	0.172139	0.134744	1.278	0.204527	p>α => Accept Null Hypothesis	Not Significant
moretime.spent	0.451503	0.116039	3.891	0.000185	***	Significant
buy.more	0.107271	0.098752	1.086	0.28011	p>α => Accept Null Hypothesis	Not Significant
design.good	-0.05349	0.160039	-0.334	0.738926	p>α => Accept Null Hypothesis	Not Significant
light.dull	0.108277	0.117792	0.919	0.360307	p>α => Accept Null Hypothesis	Not Significant
music.bothersome	-0.20886	0.104116	-2.006	0.047693	*	Not Significant
emp.not.assist	0.12882	0.101218	1.273	0.20623	p>α => Accept Null Hypothesis	Not Significant

* , **, *** -> Significant

. → Significant at 10 percent

Here, from the output the p-value = 0.000134, Hence, p-value is less than α. So, we reject the null hypothesis and accept H_1. Therefore, we can conclude that the model is statistically significant.

Step 2 : Hypothesis for ß-coefficient.

H_0 (Null Hypothesis) - All the ß-coefficient is not statistically significant.

H_1(Alternative Hypothesis) - At Least one of the ß-coefficient is statistically significant

Significant table

Coefficients:					
	Estimate	Std. Error	t value	p value	
(Intercept)	1.13789	0.31771	3.582	0.000508	***
moretime.spent	0.52414	0.09432	5.557	1.91E-07	***
emp.knowledge	-0.22314	0.10351	-2.156	0.033259	*
wall.colour	0.1865	0.08737	2.135	0.034985	*

Step 3 : Regression model

Y = α + ß1x1 + ß2x2 + ß3x3 + E

(Frequency of visit) = 1.1378 + 0.1865(wall.colour) - 0.223(Ek) + 0.524 (more time spent) + E (error term)

If employment knowledge and more time spent is constant and also if we increase the wall color by 1 unit, the frequency of visits will increase by 18.65 %. If we keep wall color and more time spent as constants and if we increase emp knowledge by 1 unit then frequency will decrease by.

After comparing the ß coefficient, we conclude that the more time spent is the more influential variable followed by employee knowledge and wall color.

Step 4 : Multicollinearity

If there is high degree positive correlation between the independent variables, then we can say there exists multi-collinearity between the variables.

moretime.spent	emp.knowledge	wall.colour
1.298247	1.313283	1.067438

Interpretation: Since the vif(Variance Inflation Factor) value for the independent variables is below 5 i.e., no multicollinearity is present between the variables

Nov 9, 2022

Sales is independent variable → Quantitative variable

Walmart → qualitative variable → (0,1)

If we have 3 variables we make dummy variable we create 2 variables

Hence n variables we have n-1 dummy variables

In walmart, column 1 represent post(walmart was open) and 0 represent pre (walmart was closed)

Sales = -332935 + 584857(PI) - 198288(WM) + 466048(FI) + 193727(Holiday) + E

Logistic Regression

Nov 14, 2022

Case -

Model: Multivariate Regression

Objective - To identify the impact of all independent variables that are **price of eggs**, **price of cookies**, on the dependent variable i.e **sales**

Justification - Since all the dependent and independent variables are numerical in nature. Hence, we will use a multivariate linear regression model.

Data Analysis:

Step 1 : Hypothesis for multivariate linear regression model

H_0 **(Null Hypothesis) -** The overall model is not statistically significant.

H_1 **(Alternative Hypothesis) -** The overall model is statistically significant.

```
Call:
lm(formula = Sales ~ Price.Eggs + Price.Cookies, data = data1)

Residuals:
   Min     1Q  Median     3Q    Max
-7.7080 -1.9511  0.3525  2.1989  6.2874

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    151.318     12.776  11.844 3.34e-12 ***
Price.Eggs     -18.727      1.882  -9.953 1.57e-10 ***
Price.Cookies   -8.786      2.369  -3.709  0.00095 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.253 on 27 degrees of freedom
Multiple R-squared:  0.8157,      Adjusted R-squared:  0.802
F-statistic: 59.75 on 2 and 27 DF,  p-value: 1.215e-10
```

Interpretation: From the output we observe that the p-value is less than 0. Hence $p < \alpha$. i.e. We reject the null hypothesis.

Step 2 : Hypothesis for ß-coefficient.

H_{0i} (Null Hypothesis) - All the ß-coefficient is not statistically significant.

H_{1i} (Alternative Hypothesis) - At Least one of the ß-coefficient is statistically significant

Coefficients:	Estimate	Std.Error	t	p value	
(Intercept)	151.318	12.776	11.844	3.34E-12	***
Price.Eggs	-18.727	1.882	-9.953	1.57E-10	***
Price.Cookies	-8.786	2.369	-3.709	0.00095	***

Interpretation: For both the variables the p value is equal to 0 i.e. is less than α. Hence we reject the null hypothesis and conclude that ß-coefficients for both the variables are statistically significant.

Step 3 : Regression model

$Y = \alpha + ß1x_1 + ß2x_2 + ß3x_3 + E$

Sales = 151.318 - 18.727(Price.Eggs) - 8.786(Price.Cookies) + E(error term)

If we increase the **price of the eggs** by 1 unit, sales will decrease 18.727 units. Similarly, if we increase the **price of the cookies** by 1 unit, sales will decrease by 8.786 units.

Step 4 : R-square (Coefficient of Determination)

Here, the R-square value is 0.8157 i.e. our model will predict the variation only 81.57 % for the dependent variable with respect to the changes in the independent variable. And the remaining 18.43% of variation is due to external factors.

Step 5 : Multicollinearity

The VIF(variance inflation factor) value for all the independent variables, therefore, there is no multicollinearity between the variables.

*A variance inflation factor (VIF) is **a measure of the amount of multicollinearity in regression analysis**. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.*

Price.Eggs	Price.Cookies
1.006466	1.006466

```
#linear regression model with the dummy variable

model2<-lm (Sales~ Price.Eggs +Price.Cookies + Ad.Type, data = data1)

model2<-lm(Sales ~ ., data = data1)

summary(model2)
```

Interpretation: Since the vif(Variance Inflation Factor) value for the independent variables is below 5 i.e., no multicollinearity is present between the variables

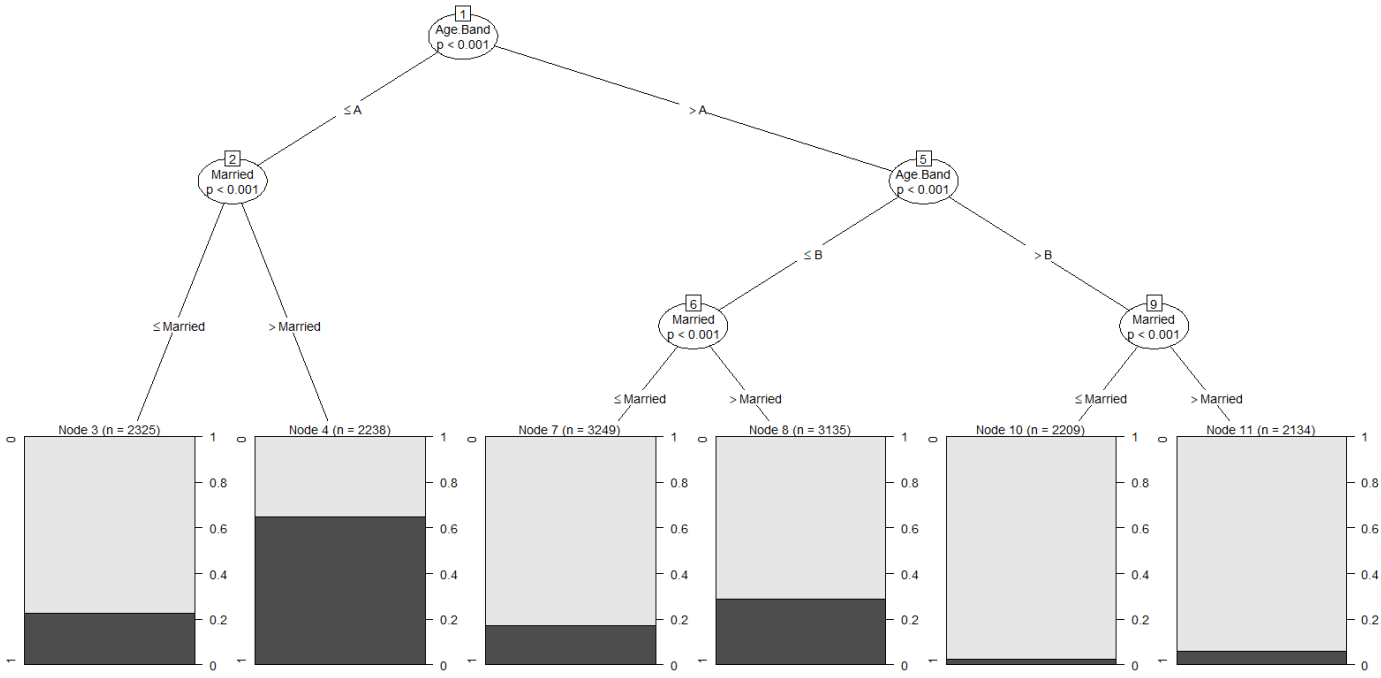
Decision Tree (CHAID Analysis)

Case Studies on Decision Tree

Case Name : Auto Insurance Policy Records

Objective - To identify the probability of non defaulter customers in a group who had paid the premium on time.

Justification - Since all the dependent qualitative in nature. Hence, we will use a CHAID(Chi Square Automatic Interaction Detection) analysis.



Interpretation:

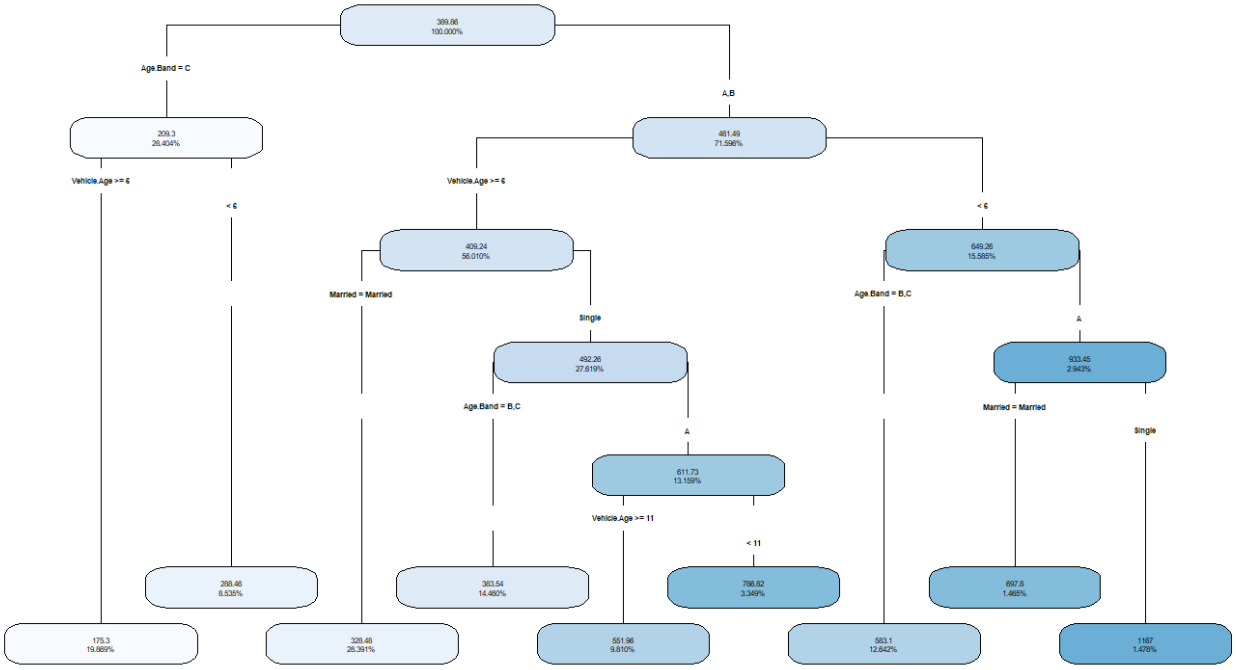
- On node no. 3, the customers belong to the age group A and they are married. Their probability of non-default is 0.2 and sample is 2325.
- On node no. 4, the customers belong to the age group A and they are single. Their probability of non-default is more than 0.6 and the sample is 2238.
- On node no. 7, the customers belong to the age group B and they are married. Their probability of non-default is about 0.2 and the sample is 3249.
- On node no. 8, the customers belong to the age group B and they are single. Their probability of non-default is more than 0.2 and the sample is 3135.
- On node no. 10, the customers belong to the age group C and they are married. Their probability of non-default is just above 0 and the sample is 2209.
- On node no. 11, the customers belong to the age group C and they are single. Their probability of non-default is about 0.1 and the sample is 2134.

CART Analysis

Objective - To identify the average losses(dependent variable) for various groups with the help decision tree.

Justification- Since the dependent variable losses is quantitative in nature, hence we will

use the CART (Classification and Regression Tree) analysis. CART analysis is based on a regression method, hence we are able to calculate average losses for various groups.



Interpretation-

- Firstly the variable age_group is going to be splitted into two branches. In one branch, we have the customers of the age group C while in the another group, the customer belongs to the age group A and B

- Furthermore, age group C customers are going to be classified with respect to the Vehicle age into two groups. One group has the vehicle more than 6 while another group has the vehicle less than 6
    - The average losses for the group (Age Band = C, vehicle age >= 6) is $175.3.
    - Similarly, the average losses for theage group c and vehicle age < 6 is $288.45
-

Logistics regression: When dependent variable is quantitative with two categories and independent variables

ClassTest

Logistics Regression

Multivariate Regression

Multivariate Regression with dummy variables

Exam

1. Market Basket Analysis
2. Multivariate Regression
3. Multivariate Regression Model
4. Logistics Regression
5. Chaid and Cart Analysis
6. RF< Analysis