Mini Project – Customer Satisfaction Analysis

By Aparajita Mukherjee

Table of Content

Project Objective

The Objective of this project is to –

1. Find evidence of multicollinearity between the variables.
2. Perform Factor Analysis by extracting four variables.
3. Naming the four factors.
4. Perform Multiple Linear Regression with Customer Satisfaction as the dependent variable and the four factors as the independent variables. Comment on model validity.

Details of the Project

A total of 12 variables were used for market segmentation in the context of product service management.

Variable	Expansion
ProdQual	Product Quality
Ecom	E-Commerce
TechSup	Technical Support
CompRes	Complaint Resolution
Advertising	Advertising
Prodline	Product Line
SalesFImage	Salesforce Image
ComPricing	Competitive Pricing
WartyClaim	Warranty & Claims
OrdBilling	Order & Billing
DepSpeed	Delivery Speed
Satisfaction	Customer Satisfaction

Required Packages

library(graphics)

library(car)

library(MASS)

library(ggplot2)

library(corrplot)

library(DataExplorer)

library(foreign)

library(lattice)

library(nFactors)

library(nortest)

library(knitr)

library(psych)

library(Hmisc)

library(scatterplot3d)

library(GPArotation)

library(factoextra)

Basic EDA(Exploratory Data Analysis)

Importing the Data

getwd()

setwd("C:/Users/HP/Documents/R Dataset")

mydata <- read.csv("GL-Factor-Hair-Revised.csv")

names(mydata)

Removing y and Irrelevant Variables

newdata <- mydata[-c(1,13)]

names(newdata)

Checking the Trend

summary(newdata)

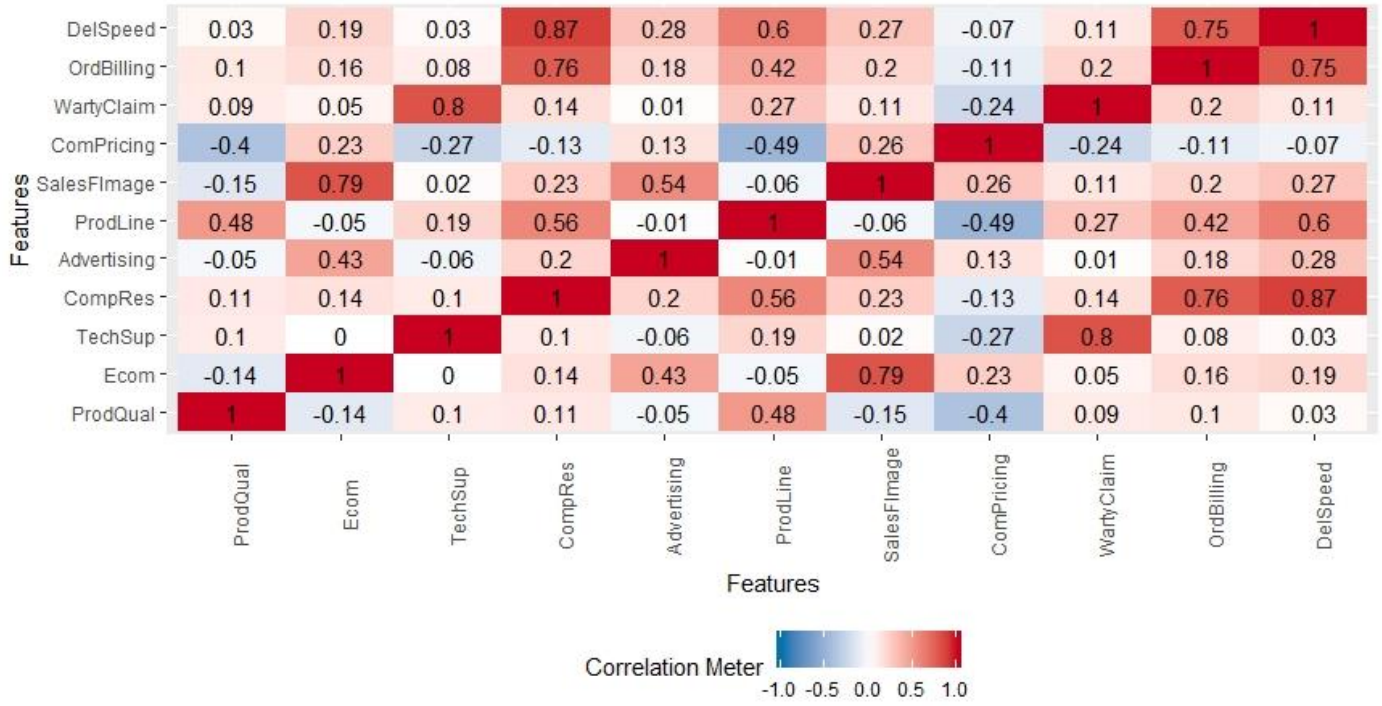
str(newdata)

dim(newdata)

Evidence of Multicollinearity

One of the assumptions of factor analysis is that there is no collinearity amongst the independent variables. But if the above statement does not hold true for a data, we would not be able to explain the effect of a particular factor on the independent variable.

plot_correlation(newdata)



The above image shows positive/negative collinearity between variables such as – DelSpeed-CompRes, WartyClaim-TechSup, OrdBilling-CompRes, etc.

Factor Analysis is therefore used to reduce the multicollinearity among independent variables. The model thus achieved after variable and collinearity reduction is further used for regression analysis.

newdatacorr <- cor(newdata)

corrplot(newdata)

cor corrplot(newdatacorr, method = 'circle',type = 'upper')

newdatacordf <- data.frame(newdatacorr)

write.csv(newdatacordf,"corrmat.csv")

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.00	-0.14	0.10	0.11	-0.05	0.48	-0.40	0.09	0.10	0.03
Ecom	-0.14	1.00	0.00	0.14	0.43	-0.05	0.23	0.05	0.16	0.19
TechSup	0.10	0.00	1.00	0.10	-0.06	0.19	-0.27	0.80	0.08	0.03
CompRes	0.11	0.14	0.10	1.00	0.20	0.56	-0.13	0.14	0.76	0.87
Advertising	-0.05	0.43	-0.06	0.20	1.00	-0.01	0.13	0.01	0.18	0.28
ProdLine	0.48	-0.05	0.19	0.56	-0.01	1.00	-0.49	0.27	0.42	0.60
ComPricing	-0.40	0.23	-0.27	-0.13	0.13	-0.49	1.00	-0.24	-0.11	-0.07
WartyClaim	0.09	0.05	0.80	0.14	0.01	0.27	-0.24	1.00	0.20	0.11
OrdBilling	0.10	0.16	0.08	0.76	0.18	0.42	-0.11	0.20	1.00	0.75
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.60	-0.07	0.11	0.75	1.00

Rcorr a Matrix of Correlations and p-value

res <- rcorr(as.matrix(newdata))

res

FlattenCorrMatrix

```
flattenCorrMatrix=function(newdatacorr,pmat){ut=upper.tri(newdatacorr)
  data.frame(
    row=rownames(newdatacorr)[row(newdatacorr)[ut]],
    row=colnames(newdatacorr)[col(newdatacorr)[ut]],
    cor=(newdatacorr)[ut],
    p=pmat[ut]  )}
```

flattencorrMatrix is the correlation matrix with row and column names as produced by cor() function

flatcorr <- flattenCorrMatrix(res1$r, res1$P)

class(flatcorr)

write.csv(flatcorr, "flatcor.csv")

Barlett's Test

print(cortest.bartlett(newdatacorr,nrow(newdata)))

$chisq
[1] 619.2726

$p.value
[1] 1.79337e-96

$df
[1] 55

The p-value is less than 0.5, therefore, dimension reduction is possible on the given dataset.

Factor Analysis(Four Factors)

Eigenvalues gives the amount of variation explained by each Principal Component(PC).

a <- eigen(newdatacorr)

eigenvalues <- a$values

eigenvectors <- a$vectors

eigenvalues

[1] 3.42697133 2.55089671 1.69097648 1.08655606 0.60942409 0.55188378 0.40151815 0.24695154 0.20355327 0.13284158

[11] 0.09842702

plot(eigenvalues, type = "lines", xlab = "Principal Components", ylab = "Eigen Value")

names(newdata)

Scree Plot

factor <- c(1,2,3,4,5,6,7,8,9,10)

scree <- data.frame(factor,eigenvalues)

plot(scree,main = "Scree Plot", col = "Blue", ylim = c(0,4), xlab = "Principal Component", ylab = "Eigen Values")

lines(scree, col = "Red")

Getting the Loadings & Communality

The analytical procedure of factor analysis involves creating uncorrelated (orthogonal) combinations of the initial independent variables. The purpose of the analysis is to reduce a mass of variables to a reasonable number of elements that the analyst can understand and explain.

principal(newdata, nfactors = length(mydata), rotate = "none")

principal(newdata, nfactors = 4, rotate = "none")

pc4_rotation <- principal(newdata, nfactors = 4, rotate = "varimax")

Using Principal Component for Scaled Dataset

newdatasc <- scale(newdata)

pr.cr <- princomp(newdatasc)

summary(pr.cr)

plot(pr.cr)

plot(pc4_rotation)

pr.cr$loadings

pr.cr$scores

pc4_rotation$scores

fviz_eig(pr.cr)

KMO Test

Kaiser-Meyer-Olkin (KMO) Test is a measure of how suited your data is for Factor Analysis. Where the KMO value -

- to 0.49 unacceptable.
- 0.50 to 0.59 miserable.
- 0.60 to 0.69 mediocre.
- 0.70 to 0.79 middling.
- 0.80 to 0.89 meritorious.
- 0.90 to 1.00 marvellous.

KMO(newdatacorr)

Kaiser-Meyer-Olkin factor adequacy

Call: KMO(r = newdatacorr)

Overall MSA = 0.65

MSA for each item =

ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling
0.51	0.63	0.52	0.79	0.78	0.62	0.62	0.75	0.51	0.76

DelSpeed

Checking the Dimension Before and After Rotation

dim(newdata)

[1] 100 11

dim(pc4_rotation$scores)

[1] 100 4

Naming Factors

Factors – RC1, RC2, RC3 & RC4, denote the amount of variation explained my each, after a "varimax" rotation.

	RC1	RC2	RC3	RC4	h2	u2	com
ProdQual	0	0.01	0.03	0.88	0.77	0.232	1
Ecom	0.06	0.87	0.05	0.12	0.78	0.223	1.1
TechSup	0.02	0.02	0.94	0.1	0.89	0.107	1
CompRes	0.93	0.12	0.05	0.09	0.88	0.119	1.1
Advertising	0.14	0.74	0.08	0.01	0.58	0.424	1.1
ProdLine	0.59	0.06	0.15	0.64	0.79	0.213	2.1
SalesFImage	0.13	0.9	0.08	0.16	0.86	0.141	1.1
ComPricing	0.09	0.23	0.25	0.72	0.64	0.359	1.5
WartyClaim	0.11	0.05	0.93	0.1	0.89	0.108	1.1
OrdBilling	0.86	0.11	0.08	0.04	0.77	0.234	1.1
DelSpeed	0.94	0.18	0	0.05	0.91	0.086	1.1

Naming the factors can be done on the basis of the group of variables explained by each. As shown below-

Process	**Promotion**	**Support**	**ProductFeatures**
RC1	**RC2**	**RC3**	**RC4**
CompRes	Ecom	TechSup	ProdQual
OrdBilling	Advertising	WartyClaim	ProdLine
DelSpeed	SalesFImage		ComPricing

```
newdf <- pc4_rotation$scores
ndata <- as.data.frame(newdf)
regPCA <- cbind(mydata$Satisfaction, ndata)
names(regPCA)[1] <- "Satisfaction"
names(regPCA)[2] <- "Process"
names(regPCA)[3] <- "Support"
names(regPCA)[4] <- "ProductFeatures"
names(regPCA)[5] <- "Promotion"
names(regPCA)
```

Multiple Linear Regression

model <- lm(regPCA$Satisfaction~regPCA$Process+regPCA$Support+regPCA$ProductFeatures+regPCA$Promotion, data = regPCA)

summary(model)

anova(model)

confint(model)

Model Validity

Linear Model

model <- lm(regPCA$Satisfaction~regPCA$Process+regPCA$Support+regPCA$ProductFeatures+regPCA$Promotion, data = regPCA)

summary(model)

```
Call:

lm(formula = regPCA$Satisfaction ~ regPCA$Process + regPCA$Support +

    regPCA$ProductFeatures + regPCA$Promotion, data = regPCA)


Residuals:

     Min       1Q   Median       3Q      Max

-1.6308 -0.4996  0.1372  0.4623  1.5228


Coefficients:

                          Estimate   Std. Error  t value   Pr(>|t|)

(Intercept)               6.91800    0.07089     97.589    < 2e-16 ***

regPCA$Process            0.61805    0.07125     8.675     1.12e-13 ***

regPCA$Support            0.06714    0.07125     0.942     0.348

regPCA$ProductFeatures    0.54032    0.07125     7.584     2.24e-11 ***

regPCA$Promotion          0.50973    0.07125     7.155     1.74e-10 ***

---

Signif. codes:          0 ‘***’    0.001 ‘**’    0.01 ‘*’    0.05 ‘.’    0.1 ‘ ’   1
```

Residual standard error: 0.7089 on 95 degrees of freedom

Multiple R-squared: 0.6605, Adjusted R-squared: 0.6462

F-statistic: 46.21 on 4 and 95 DF, p-value: < 2.2e-16

The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value

- In our model, it can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. The p-value is much lower than alpha of 5% level, we will therefore reject the null hypothesis(all betas = 0). Hence, concluding that, at least, one of the predictor variables is significantly related to the outcome variable.
- Satisfaction($\hat{y}$) = 6.9+0.61*Process+0.06*Support+0.54*ProductFeatures+0.51*Promotion.
- The above output/model proves that there is a strong positive relationship of customer satisfaction with Process, ProductFeatures and Promotion.
- Adjusted R-squared is the Variance explained by the model. In your model, the model explained 64 percent of the variance of y. R squared is always between 0 and 1. The higher the better.

ANOVA

We can use the ANOVA test to estimate the effect of each feature on the variances with the anova() function.

anova(model)

Analysis of Variance Table

Response: regPCA$Satisfaction

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
regPCA$Process	1	37.816	37.816	75.253	1.118e-13 ***
regPCA$Support	1	25.723	25.723	51.188	1.740e-10 ***
regPCA$ProductFeatures	1	0.446	0.446	0.888	0.3484
regPCA$Promotion	1	28.903	28.903	57.515	2.245e-11 ***
Residuals	95	47.740	0.503		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Confidence Interval

confint(model)

	2.5 %	97.5 %
(Intercept)	6.77726797	7.0587320
regPCA$Process	0.47660590	0.7594879
regPCA$Support	0.36829308	0.6511751
regPCA$ProductFeatures	-0.07430515	0.2085769
regPCA$Promotion	0.39887803	0.6817601

Source of Data

- Great Learning Mentored Learning Session PCA & FA
- Great Learning Multiple Linear Regression 1,2 & 3 recorded sessions.
- Google
- R Blogger