**SUMMARY**

Here we are provided with the data set **‘Marketing-Customer-Value-Analysis’.**

Data source: IVY Professional Business School.

Given the data set, we are asked to predict the **Customer Lifetime Value.** As we know for that we have to divide the data set into target variable and predictor variable. Target variable are those variables which we have to predict whereas Predictors are the independent variables which are use to predict the target variable.

**Approach:**

Given the dataset, we found that our target variable is “Customer.Lifetime.Value” which is continuous in nature. In other words, “Customer.Lifetime.Value” consists of numeric values.

Since the target variable is continuous in nature, therefore we followed **Linear Regression Model** for our further approach.

Our data set consists of both continuous as well as categorical columns. Before we discussed about what is the meaning of continuous columns, categorical columns are those columns which are predefined, i.e. for example , yes or no, gender type etc.

Here, our target variable is continuous in nature and some of the predictors are continuous in nature so for visual relationship we used *histogram*, *scatter* *plot* and for testing the significance of those continuous predictors we use **Correlation.**

Similarly, for categorical predictors we used *barplot* and *boxplots* for visual relation and for testing the significance of the column we use **Anova test.**

We used *Variance Inflation Analysis* for mulcolinearity test, *Bruch-Pegan test* for homoscedasticity test, *Durbin-Watson test* for autocorrelation and *Anderson-Darling test* for normality.

**Steps & Interpretations:**

**Step 1:**

#importing data to Rstudio:

CustomerLifetimeValue=read.csv('C:/Users/LENOVO/Downloads/Final R Project IVY/Final R Project IVY/Fn-UseC\_-Marketing-Customer-Value-Analysis.csv',na.string=c(""," ","NA","NULL"), stringsAsFactors = T)

#str command helps us to give the overall summary of the imported columns in R

str(CustomerLifetimeValue)

**Step 2:**

#converting few columns inside the data as a categorical column

CustomerLifetimeValue$Type.of.Open.Complaints=as.factor(CustomerLifetimeValue$Type.of.Open.Complaints)

CustomerLifetimeValue$Type.of.Policies=as.factor(CustomerLifetimeValue$Type.of.Policies)

#table command gives us the description of the column whether it has unique values or not

table(CustomerLifetimeValue$Type.of.Open.Complaints)

table(CustomerLifetimeValue$Type.of.Policies)

str(CustomerLifetimeValue)

**Step 3**

#treatment of the useless columns. We used null to remove the garbage or the useless column from the data set

UselessColumns=c('Customer','Effective.To.Date')

CustomerLifetimeValue[,UselessColumns]=NULL

#Head command is used to give us the first 6 rows of all the column. Here 6 is the default value, if we want more rows we can specify that is the code

head(CustomerLifetimeValue)

str(CustomerLifetimeValue)

**Step 4:**

#outliers treatment. Here we used boxplots to have a look on which column we have outliers.

boxplot(CustomerLifetimeValue$Income, horizontal = T)

boxplot(CustomerLifetimeValue$Monthly.Premium.Auto, horizontal = T) #outliers

boxplot(CustomerLifetimeValue$Months.Since.Last.Claim, horizontal = T)

boxplot(CustomerLifetimeValue$Months.Since.Policy.Inception, horizontal = T)

boxplot(CustomerLifetimeValue$Total.Claim.Amount, horizontal = T) #outliers

#Treating outliers: Here we used quantile for treating the outliers. So that values which are far away can come close to eachother.

####Treatment for Monthly.Premium.Auto

qt=quantile(CustomerLifetimeValue$Monthly.Premium.Auto,c(0.95,0.96,0.963,0.965,0.97,0.98,0.99,0.995,0.996,0.997,0.998,0.999),na.rm = TRUE)

qt

qt\_final=quantile(CustomerLifetimeValue$Monthly.Premium.Auto,0.995,na.rm = TRUE)

qt\_final[1]

max(CustomerLifetimeValue$Monthly.Premium.Auto)

CustomerLifetimeValue[,"Monthly.Premium.Auto"] = ifelse(CustomerLifetimeValue[,"Monthly.Premium.Auto"] > qt\_final[1] , qt\_final[1], CustomerLifetimeValue[,"Monthly.Premium.Auto"])

boxplot(CustomerLifetimeValue$Monthly.Premium.Auto, horizontal = T)

max(CustomerLifetimeValue$Monthly.Premium.Auto)

####Treatment for Total.Claim.Amount

qt2=quantile(CustomerLifetimeValue$Total.Claim.Amount,c(0.95,0.96,0.963,0.965,0.97,0.98,0.99,0.995,0.996,0.997,0.998,0.999),na.rm = TRUE)

qt2

qt\_final2=quantile(CustomerLifetimeValue$Total.Claim.Amount,0.995,na.rm = TRUE)

qt\_final2[1]

max(CustomerLifetimeValue$Total.Claim.Amount)

CustomerLifetimeValue[,"Total.Claim.Amount"] = ifelse(CustomerLifetimeValue[,"Total.Claim.Amount"] > qt\_final2[1] , qt\_final2[1], CustomerLifetimeValue[,"Total.Claim.Amount"])

boxplot(CustomerLifetimeValue$Total.Claim.Amount, horizontal = T)

max(CustomerLifetimeValue$Total.Claim.Amount)

**Step 5:**

#exploring the multiple continuous features. Here we separated the continuous columns

ColsForHist=c('Customer.Lifetime.Value','Income','Monthly.Premium.Auto','Months.Since.Last.Claim','Months.Since.Policy.Inception','Total.Claim.Amount')

#splitting the windows: Since there are 6 column so we divide the window into 6 parts.

par(mfrow=c(2,3))

#for color palatte: we installed this library for the colour palate

library(RColorBrewer)

#for loop function: we used loop function for histogram for all the continuous columns togethers in a window.

for (hist\_cols in ColsForHist) {

hist(table(CustomerLifetimeValue[,c(hist\_cols)]),main = paste("Histogram of:",hist\_cols),col=brewer.pal(8,"Paired"))

}

**Step 6:**

#exploring the multiple categorical features: as similar as continuous columns.

ColsForBar1=c('State','Response','Coverage','Education','EmploymentStatus','Gender','Location.Code')

ColsForBar2=c('Marital.Status','Type.of.Open.Complaints','Type.of.Policies','Policy.Type','Policy','Renew.Offer.Type','Sales.Channel','Vehicle.Class','Vehicle.Size')

#splitting the window for set 1

par(mfrow=c(2,4))

#for using color palatte

library(RColorBrewer)

#loop function foe set 1

for (bar\_cols in ColsForBar1) {

barplot(table(CustomerLifetimeValue[,c(bar\_cols)]),main = paste("Barplot of :", bar\_cols), col = brewer.pal(8,"Paired"))

}

#splitting the window for set 2

par(mfrow=c(2,5))

#loop function for set 2

for (bar\_cols in ColsForBar2) {

barplot(table(CustomerLifetimeValue[,c(bar\_cols)]),main = paste("Barplot of :", bar\_cols), col = brewer.pal(8,"Paired"))

}

**Step 7:**

#Statistical relationship between target veriable and the predictor

## continuous vs continuous : Correlation Test

## continuous vs categorical : Anova Test

#Visual Relationship

## Continuous vs Continuous : Scatter plot

## Continuous vs Categorical : Bar Plot

# Continuous vs Continuous : Scatter plot.

##For multiple columns at once

ContinuousCols= c('Customer.Lifetime.Value','Income','Monthly.Premium.Auto','Months.Since.Last.Claim','Months.Since.Policy.Inception','Total.Claim.Amount')

plot(CustomerLifetimeValue[,ContinuousCols],col='blue')

# Correlation test: to measure the significance of the columns

## for multiple columns

CorData= cor(CustomerLifetimeValue[,ContinuousCols],use='complete.obs')

class(CorData)

CorData[,'Customer.Lifetime.Value']

names(CorData[,'Customer.Lifetime.Value'])

#setting the threshold level

abs(CorData[,'Customer.Lifetime.Value'])>0.5

# as we can see we cannot include any of the columns, so we decreased the threshold level

abs(CorData[,'Customer.Lifetime.Value'])>0.2

names(CorData[,'Customer.Lifetime.Value'][abs(CorData[,'Customer.Lifetime.Value'])>0.2])

## so here we sorted out the name of those columns which satisfied the threshold level.

#Continuous Vs Categorical----- Box Plot

catVar=c('State','Response','Coverage','Education','EmploymentStatus','Gender','Location.Code','Marital.Status','Type.of.Open.Complaints','Type.of.Policies','Policy.Type','Policy','Renew.Offer.Type','Sales.Channel','Vehicle.Class','Vehicle.Size')

for(bar\_cols in catVar ){

boxplot(Customer.Lifetime.Value ~ (CustomerLifetimeValue[,c(bar\_cols)]), data = CustomerLifetimeValue, main = paste('Boxplot of:',bar\_cols), col = brewer.pal(8,"Paired"))

}

## Anova Test: to test the significance level of the categorical columns.

###Null Hypothesis: the variables are not correlated

#creating a loop

for(i in catVar){

test\_summary= summary(aov(Customer.Lifetime.Value ~ CustomerLifetimeValue[,c(i)], data = CustomerLifetimeValue))

print(paste("Anova test of :",i))

print(test\_summary)

}

**Step 8:**

#creating a separate database: for better interpretation of the data, after treating all the columns.

ImpData= CustomerLifetimeValue

TargetVariableName= 'Customer.Lifetime.Value'

BestPredictorName= c('Monthly.Premium.Auto','Total.Claim.Amount','Coverage','Education','EmploymentStatus','Marital.Status','Type.of.Open.Complaints','Type.of.Policies','Renew.Offer.Type','Vehicle.Class','Vehicle.Size')

TargetVariable= ImpData[,c(TargetVariableName)]

PredictorVariable=ImpData[,BestPredictorName]

str(PredictorVariable)

#creating final data for ML

DataForML=data.frame(TargetVariable,PredictorVariable)

str(DataForML)

head(DataForML)

#splitting data into 70% for training and 30% for testing

TrainingSampleIndex= sample(1:nrow(DataForML),size = 0.7\*nrow(DataForML))

#length command gives us the total length of the data

length(TrainingSampleIndex)

#to get the records of the training set

DataForMLTrain=DataForML[TrainingSampleIndex,]

DataForMLTrain

#index for test set

DataForMLTest=DataForML[-TrainingSampleIndex,]

DataForMLTest

dim(DataForMLTrain)

dim(DataForMLTest)

**Step 9:**

#linear regression algorithm

StartTime=Sys.time()

Model\_Reg=lm(TargetVariable~.,data= DataForMLTrain)

EndTime= Sys.time()

#we used this to check how much time was required to run the algorithm

TimeTaken= EndTime-StartTime

summary(Model\_Reg)

#multicolinearity test: used to manipulate those columns which are highly correlated with the other independent columns.

library(car)

Multi=vif(Model\_Reg)

last\_col=Multi[,3]

FinalValues=last\_col\*\*2

Multi\_Final=data.frame(FinalValues)

Multi\_Final

#removal of column: after treating the multicollinearity test we removed the column which is no more in use.

Model\_Reg=lm(TargetVariable~.-Monthly.Premium.Auto,data=DataForMLTrain)

Model\_Reg

summary(Model\_Reg)

#homoscadasticity test

library(lmtest)

bptest(Model\_Reg)

#Durbin\_Watson test: to check the autocorrelation part

dwtest(Model\_Reg)

#Anderson\_Darling Normality test: to chech whether the errors are normally distributed or not.

library(nortest)

ad.test(Model\_Reg$residuals)

**Step 10:**

#based on the model using train data, we will predict values on test data

DataForMLTest$Pred\_LM=predict(Model\_Reg, DataForMLTest)

head(DataForMLTest)

#calculating the absolute error % for each prediction

DataForMLTest$LM\_APE=100\*(abs(DataForMLTest$TargetVariable-DataForMLTest$Pred\_LM)/DataForMLTest$TargetVariable)

head(DataForMLTest)

#checking the accuracy of the model on test data using both mean and median

print(paste('Mean Accuracy of Linear Regression Model is: ', 100 - mean(DataForMLTest$LM\_APE)))

print(paste('Median Accuracy of Linear Regression Model is: ', 100 - median(DataForMLTest$LM\_APE)))

**Result Found:**

After predicting the model, we got,

Here in our script we performed different tests and we found there results. Now the results may vary from person to person because every time we run the model, it takes different sets of values and provide us with different results.

Correlation test- In the beginning we performed the correlation test, which implies whether the predictor(continuous) do have any correlation with the target variable (continuous) or not and for that we have to set the threshold level. We know correlation lies between -1 to 1 and by looking into the data we set out threshold level to 0.2 where only monthly premium auto and total claim amount column satisfied the level. So from this we can conclude that with our target variable this two column do have some relation.

Anova test- Next we performed this test which gives us the relationship between target variable(continuous) and predictor(categorical). Here the null hypothesis implies that the variables are not correlated. So here we got, coverage, education, employment status, marital status, type of open complaint, type of polies, renewal offer type, vehicle class, vehicle size are the columns that reject the null hypothesis and are having some relation with the target variable.

Multicollinearity test- This implies one variable have correlation with other variables. So if the variable lies below 1 then it has no multicollinearity, if it lies between 1- 5 then it is having moderate multicollinearity and if it is more that 5 then high multicollinearity is present. And here monthly premium auto is having multicollinearity, so removed that column from the model.

Next we fitted the linear regression model, and we got the value of r-square as 0.6191 and adjusted r-square as 0.6169.

Next we performed, Anderson Darling test which helps us to check whether the errors are normally distributed or not. And here we see, that it is rejecting the null hypothesis i.e. below 0.05. Therefore errors are not normally distributed.

Durbin-Watson Test helps us to check whether there is any autocorrelation or not. Autocorrelation implies whether the residuals are not independent from each other. Here we can see that we accept the null hypothesis i.e. there is no autocorrelation.

Breusch-Pegan Test helps us to check whether the error variance are equal or not. After performing the test we conclude that we reject the null hypothesis and can say that error variance are not equal i.e. p-value is below 0.05.

From this two set of data, we calculate the accuracy level i.e.

Mean Accuracy of Linear Regression Model is: 73.33, rounded upto 2 decimal

places.

And Median Accuracy of Linear Regression Model is:

81.95 rounded upto 2 decimal places.

**Significance of the Variables:**

Given our dataset, except our target variable, there are 21 variables. But

including the target variable, we got 22 variables.

Out of this 21 predictors, some of the variables are significant and some of

them are not.

First of all, by looking into the data, we would try to relate the predictors with

the target variable and try to detect whether it is a garbage column or not.

So, here by looking at the dataset we concluded that, ‘Customer’ and

‘Effective.To.Date’ are the garbage column and we can say it is an insignificant column.

Next, we look into the correlation part. There we considered the continuous predictors. According to our set threshold level, i.e., 0.2, there are two columns which are significant. They are ‘Monthly.Premium.Auto’ and ‘Total.Claim. Amount’.

Similarly by looking into the Anova test, we got 'Coverage','Education','EmploymentStatus','Marital.Status','Type.of.Open.Complaints','Type.of.Policies','Renew.Offer.Type','Vehicle.Class','Vehicle.Size' are the significant columns.

After performing the multicollinearity test, from the model summary finally we got:

Coverage with 0.1% significance level.

Employment status (Employed, Medical Leave) with 0.1% significance level &

10% significance level.

Type of open complaint where type 3 with 5% and type 4 with 1% significance

Level.

Type of policies with 0.1% significance level.

Vehicle class with 0.1% significance level.

Now for different trials the significance level of different columns changes as at each trial different sets of values are taken into consideration. So here the significant variables are taken into consideration as maximum number of times this variables became the significant one.

**Business Interpretation:**

Here our target variable is Customer Lifetime value which implies the total

revenue of the company from their entire relationship with the customers, i.e.

total amount of money customer is expected to spend in one’s company.

Now let us check the relationship between the target variable and the predictors and interpret there business result.

So let us first start with the customer column, it is not such useful for the

business purpose. Because every time a new customer will add , a new id will

generate, so customer id is of no use.

State, says where the customer belongs to. This column again not much useful.

Because customers can stay at different places and it does not reveal about their capacity of investment and nevertheless we found out that it is not correlated

with the target variable after using anova test.

Response means responsive to different companies courtesis. After doing the

test it is again not much important column. Because people may have not much habituated with the technology or they may not have the proper access to know about different schemes, so it is not much useful.

Coverage says the different coverage schemes by the company. It is an

important column which gives us an idea of different schemes for different

customers the companies want to cover.

Education level is again an important part because those who have higher

education can have proper knowledge and earning power to invest in companies.

Effective to date of the scheme is a garbage column which does not create any impact to business.

Employment Status is an important column because unemployed or retired or disabled don’t have the capacity to invest whereas others have the capacity to invest.

Gender does not give much importance to the business because now-a-days

both the genders are equally successful in every sector.

Therefore like this way we can say that,

|  |  |
| --- | --- |
| Income | |
| Location Code   |  | | --- | | Monthly Premium Auto | | Months Since Last Claim | | Months Since Policy Inception | | Policy Type | | Policy |   Sales Channel  This are the above few columns which are not much important for the  Company.  Whereas marital status, type of open complaints, type of policies, renewal  offer type, vehicle class and type are important to the company.  Because type of open complaint help the company know about the type of  complaints customers are doing. Different polices helps the customers to  know about the different contents of the policies. Renewal offer type helps  the company to know about the customers intensions about the offers. And vehicle  Size and type tells about different type and size of the vehicle customers are looking  For. | |
|  |
|  |