R – Project: “Internet Data “

Course: “Data Science with R”

Business Scenario:

Description:Analyse the internet data of www.datadb.com

The web analytics team of www.datadb.com is interested to understand the web activities of the site, which are the sources used to access the website. They have a database that states the keywords of time in page, source group, bounces, exits, unique page views, and visits. The variables in the dataset are defined here for better understanding:

**- Bounces:** It represents the percentage of visitors who enter the site and "bounce" (leave the site) rather than continuing to view other pages within the same site.

**- Exits:** It represents the percentage of visitors to a site who actively click away to a different site from a specific page, after possibly having visited any other page on the site.

**- Continent:** It shows the continent from which the site has been accessed.

**- Source group:** It shows how the visitor has accessed the site.

**- Time on page:** It shows how long the user has spent on that particular page of the website.

- **Unique page view:** It represents the number of sessions during which that page was viewed one or more times.

**- Visits:** A visit counts all visitors, no matter how many times the same visitor may have been to your site.

`

Expectation /Goals:

The team is targeting at the following issues:

The team wants to analyse each variable of the data collected through data summarization to get a basic understanding of the dataset and to prepare for further analysis.

As mentioned earlier, a unique page view represents the number of sessions during which that page was viewed one or more times. A visit counts all instances, no matter how many times the same visitor may have been to your site. So the team needs to know whether the unique page view value depends on visits.

Find out the probable factors from the dataset, which could affect the exits. Exit Page Analysis is usually required to get an idea about why a user leaves the website for a session and moves on to another one. Please keep in mind that exits should not be confused with bounces.

Every site wants to increase the time on page for a visitor. This increases the chances of the visitor understanding the site content better and hence there are more chances of a transaction taking place. Find the variables which possibly have an effect on the time on page.

A high bounce rate is a cause of alarm for websites which depend on visitor engagement. Help the team in determining the factors that are impacting the bounce.

R-Code File:



R-Code with Output Details and Analysis Details:

# Title: "Internet Data"

# Author: "SANJAY TALLOLLI"

#---------------------------------------------------------------------------

getwd()

setwd("I:\\SIMPLILEARN COURSES LIVE 2018\\DATASCIENCE WITH R\\COURSE MATERIALS\\Project\\Projects for Submission\\Internet")

#Goal / Expectation (1):

#The team wants to analyse each variable of the data collected through data summarization

#to get a basic understanding of the dataset and to prepare for further analysis.

mydata<-read.csv("I:\\SIMPLILEARN COURSES LIVE 2018\\DATASCIENCE WITH R\\COURSE MATERIALS\\Project\\Projects for Submission\\Internet\\Internet\_Dataset1.csv",header = T)

summary(mydata)

**# OUTPUT:**

## Bounces Exits Continent

## Min. : 0.000 Min. : 0.000 AF : 321

## 1st Qu.: 0.000 1st Qu.: 1.000 AS : 3171

## Median : 1.000 Median : 1.000 EU : 6470

## Mean : 0.713 Mean : 0.906 N.America:20043

## 3rd Qu.: 1.000 3rd Qu.: 1.000 OC : 1356

## Max. :30.000 Max. :36.000 SA : 748

##

## Sourcegroup Timeinpage Uniquepageviews

## google :11542 Min. : 0.00 Min. : 1.000

## (direct) : 7532 1st Qu.: 0.00 1st Qu.: 1.000

## Others : 5360 Median : 0.00 Median : 1.000

## tableausoftware.com : 2388 Mean : 73.18 Mean : 1.114

## t.co : 2249 3rd Qu.: 10.00 3rd Qu.: 1.000

## public.tableausoftware.com: 1354 Max. :46745.00 Max. :45.000

## (Other) : 1684

## Visits BouncesNew

## Min. : 0.000 Min. :0.00000

## 1st Qu.: 1.000 1st Qu.:0.00000

## Median : 1.000 Median :0.01000

## Mean : 0.906 Mean :0.00713

## 3rd Qu.: 1.000 3rd Qu.:0.01000

## Max. :45.000 Max. :0.30000

##

**#ANALYSIS:**

#bounces min=0,max=30

#exit min=0 max=36

#From the result of summarized dataset, it is observed that the numerical data includes

#information related to the maximum, minimum, and mean data.

#The categorical data like continent includes the data of the number of times the category has been

#repeated in the dataset. We can see that there is a maximum value of 30 bounces for the website.

#This site was accessed maximum number of times by visitors from North A

#Goal / Expectation (2):

#As mentioned earlier, a unique page view represents the number of sessions during which that page was viewed one or more times. A visit counts all instances, no matter how many times the same visitor may have been to your site. So the team needs to know whether the unique page view value depends on visits.

cor(mydata$Uniquepageviews,mydata$Visits)

**# OUTPUT:**

## [1] 0.8144457

ano<-aov(Uniquepageviews~Visits, data=mydata)

summary(ano)

**# OUTPUT:**

## Df Sum Sq Mean Sq F value Pr(>F)

## Visits 1 8052 8052 63257 <2e-16 \*\*\*

## Residuals 32107 4087 0

## ---

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

**#ANALYSIS:**

#We can infer from the results that the visits variable has a significant impact on

#Unique.Pageviews. So the team can conclude that unique page values depend on visits.

#Goal / Expectation (3):

#Find out the probable factors from the dataset, which could affect the exits.

#Exit Page Analysis is usually required to get an idea about why a user leaves the

#website for a session and moves on to another one. Please keep in mind that exits should

#not be confused with bounces

anoo<-aov(Exits~.,data = mydata)

summary(anoo)

**# OUTPUT:**

## Df Sum Sq Mean Sq F value Pr(>F)

## Bounces 1 10578 10578 1.043e+05 < 2e-16 \*\*\*

## Continent 5 3 1 5.960e+00 1.62e-05 \*\*\*

## Sourcegroup 8 7 1 8.760e+00 4.89e-12 \*\*\*

## Timeinpage 1 130 130 1.279e+03 < 2e-16 \*\*\*

## Uniquepageviews 1 1573 1573 1.552e+04 < 2e-16 \*\*\*

## Visits 1 1 1 5.014e+00 0.0251 \*

## Residuals 32091 3254 0

## ---

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

**#ANALYSIS:**

#From the result of ANOVA given here, we can see that source.group, bounces,

#and unique.pageviews have more significance. Visits have comparatively less significance.

#Hence we can say that exit from the site is affected by the factors of source group,

#bounces, and unique.pageviews.

#Goal / Expectation (4):

#Every site wants to increase the time on page for a visitor.

#This increases the chances of the visitor understanding the site content better and

#hence there are more chances of a transaction taking place.

#Find the variables which possibly have an effect on the time on page.

anooo<-aov(Timeinpage~.,data = mydata)

summary(anooo)

**# OUTPUT:**

## Df Sum Sq Mean Sq F value Pr(>F)

## Bounces 1 5.947e+07 59466495 422.868 < 2e-16 \*\*\*

## Exits 1 1.304e+08 130400662 927.283 < 2e-16 \*\*\*

## Continent 5 4.767e+06 953431 6.780 2.51e-06 \*\*\*

## Sourcegroup 8 1.545e+06 193153 1.374 0.202

## Uniquepageviews 1 1.791e+08 179133934 1273.826 < 2e-16 \*\*\*

## Visits 1 1.073e+08 107321113 763.163 < 2e-16 \*\*\*

## Residuals 32091 4.513e+09 140627

## ---

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

**#ANALYSIS:**

#only source group is not affecting the time in page views rest all are significantly affecting the time-in page views

#Goal / Expectation (5):

#A high bounce rate is a cause of alarm for websites which depend on visitor engagement.

#Help the team in determining the factors that are impacting the bounce

#this bounce rate is having variables

#data for the variable bounces has to be between 0 and 1,

mydata$Bounces=mydata$Bounces\*0.01

rmm<-glm(Bounces~Timeinpage+Continent+Exits+Sourcegroup+Uniquepageviews+Visits,data = mydata,family = "binomial")

**# OUTPUT:**

## Warning in eval(family$initialize): non-integer #successes in a binomial

## glm!

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(rmm)

**# OUTPUT:**

## Call:

## glm(formula = Bounces ~ Timeinpage + Continent + Exits + Sourcegroup +

## Uniquepageviews + Visits, family = "binomial", data = mydata)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.26149 -0.02406 0.00206 0.00895 1.81288

##

## Coefficients:

## Estimate Std. Error z value

## (Intercept) -4.9667681 0.6784678 -7.321

## Timeinpage -0.0010294 0.0005774 -1.783

## ContinentAS 0.0022768 0.6932044 0.003

## ContinentEU -0.0069240 0.6786600 -0.010

## ContinentN.America 0.0101334 0.6674188 0.015

## ContinentOC 0.0201123 0.7333671 0.027

## ContinentSA 0.0237507 0.7914250 0.030

## Exits 1.3907608 0.3356504 4.143

## Sourcegroupfacebook -0.0241949 1.1045171 -0.022

## Sourcegroupgoogle -0.0783631 0.1720157 -0.456

## SourcegroupOthers -0.0767919 0.2182692 -0.352

## Sourcegrouppublic.tableausoftware.com -0.2528285 0.4923123 -0.514

## Sourcegroupreddit.com -0.0092792 0.4709304 -0.020

## Sourcegroupt.co 0.0148690 0.2760157 0.054

## Sourcegrouptableausoftware.com -0.1129305 0.3190762 -0.354

## Sourcegroupvisualisingdata.com -0.0822525 0.4614866 -0.178

## Uniquepageviews -3.2363108 0.5791664 -5.588

## Visits 2.1941121 0.5202216 4.218

## Pr(>|z|)

## (Intercept) 2.47e-13 \*\*\*

## Timeinpage 0.0746 .

## ContinentAS 0.9974

## ContinentEU 0.9919

## ContinentN.America 0.9879

## ContinentOC 0.9781

## ContinentSA 0.9761

## Exits 3.42e-05 \*\*\*

## Sourcegroupfacebook 0.9825

## Sourcegroupgoogle 0.6487

## SourcegroupOthers 0.7250

## Sourcegrouppublic.tableausoftware.com 0.6076

## Sourcegroupreddit.com 0.9843

## Sourcegroupt.co 0.9570

## Sourcegrouptableausoftware.com 0.7234

## Sourcegroupvisualisingdata.com 0.8585

## Uniquepageviews 2.30e-08 \*\*\*

## Visits 2.47e-05 \*\*\*

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 234.937 on 32108 degrees of freedom

## Residual deviance: 96.514 on 32091 degrees of freedom

## AIC: 506.56

##

## Number of Fisher Scoring iterations: 11

**#ANALYSIS:**

#As can be inferred from the result shown, the BouncesNew, Unique.Pageviews and visits are the variables that impact the target variable bounces. It has greater significance.

========================== END ===============================