A Regression Model for Analysis of Bounce Rate Using Web Analytics

Meenakshi Mittal, S. Veena Dhari

Abstract*: *Bounce rate is an effective parameter to measure the quality of any website. Bounce rate refers to the percentage of visitors that leave a website (or "bounce" back to the search results or referring website) after viewing only one page a website. High bounce rate is bad as it depicts that the content on a site didn't match what the visitor was looking for so he left without viewing another page. Since bounce rate equates to visitors taking absolutely no action on a website so this metric could be used as a measure of success .This paper analyses the bounce rate of a website based on web analytics data. In this paper, analysis of bounce rate will be based on performance of website. Data is collected using Google Analytics tool. After applying preprocessing techniques to data an eleven step regression model is built using the various attributes like Average Server Response Time, Average Server Connection Time, Average Redirection Time, Average Page Download Time, Average Domain Lookup time and Average Page Load Time. Mathematical equation is constructed on the basis of outcome of result so that bounce rate can be analyzed and predicted. Model is further refined after establishing the correlation between various attributes. Correlation is established to improve the accuracy in analysis and prediction of bounce rate. This regression model gives insight about the various parameters involved and their effect on bounce rate. Qunatile Quantile plot is constructed to see if plausible data is normally distributed. This complete experiment is done using R Studio

Index Terms*: *Bounce Rate, Google Analytics, Regression model, Web Analytics, Website Performance

I. INTRODUCTION

Highlight Bounce rate is the percentage of single-page sessions on a website, meaning a visitor left website from an entrance page without visiting another page on website. A websites trust should be measured as decision makers may need the information for supplementing their decision-making in symptom analysis using information from health related websites. TNM (Trust Necessitated through Metrics) is used for measuring the quality of website. For measuring the TRUST a tool is used which is based on various parameters like Average time on website, Bounce rate, Average daily visits, Page/visit and Category rank.[1] There may be many tools used for collecting web analytics data like Google Analytics , Stat counter and server log files. [2]An average bounce rate can be within the range of 41-60% but an excellent bounce rate will be between 20-40%. Generally majority of websites fall into the 41-60% range. A website or blog having bounce rate more then 70% needs improvement and should pay attention to website content, structure and page load speed. A high bounce rate can be a sign that content of website is not targeted to the visitor's requirements.

On the other hand high bounce rate can also be used a positive indicator depending on the nature of business. Most marketing experts believe that bounce rate may affect the search engine rankings of a website. Bounce rate may impact search Engine rankings, either directly or indirectly. [3] Google's Gary Illyes stated in 2015 that Google do not use bounce rate data from Google Analytics for search engine rankings. So a website may not be directly affected by bounce rate in search rankings, but it does not mean that performance of website can be termed as best. Bounce rate's affects website's goals. If webmasters are looking to convert leads on your website, bounce rate may play very important to company's site. [4]- [5]
Some of the reasons for high bounce rate-

1) For a single page website bounce rate will always be 100%.
2) Google Analytics [6] tracking code may not be inserted correctly. If the code is inserted in header or footer then also bounce rate will be 100%
3) Coding errors can also affect bounce rate. Bounce rate is determined based on cookies, and if coding is resetting that cookie, it resets the session and sends inaccurate data to your Google Analytics account.
4) If a website is poorly designed then user neither will nor is able to find what it is looking for. So a visitor may leave the website quickly adding to the bounce rate of website.
5) Poor Search Engine Optimization may be another factor for high bounce rate. If wrong keyword strategy is used for ranking a website in search engine then visitors may not find required information and may leave website.
6) If a user bookmarks a page, visits that page and leaves, it's still considered a bounce.[7]

This paper is organized as follows - Section I contains the introduction about the bounce rate, various factors responsible for bounce rate, standard bounce rates for different types of websites and how bounce rate affects the websites performance . Section II contains the previous work done to measure and reduce the bounce rate of website. Section III describes the methodology that is used to create and refine the regression model where as Section IV contains results and discussion about the model where as section V concludes the study.

II. RELATED WORK

In [7] author talks about

Revised Manuscript Received on June 15, 2019.

Meenakshi Mittal , Department of Computer Application, Rabindranath Tagore University, Bhopal, India
S Veena Dhari , Department of Computer Science and Engineering, Rabindranath Tagore University, Bhopal, India

creating a conceptual model for calculating TRUST based on data obtained from web analytics for content driven sites. Here many factors like Dwell Time, over all traffic bounce rate, conversion rate and website quality are being used to create the model.

In [8] author uses web analytics for segmentation of B2B websites. Here Bounce rate for registered and unregistered users on a website is studied. Considering B2B characteristics in analytics, effectiveness of user behavior by segmentation was studied.

In [9] author compares various tools like Google analytics , stat counter to collect data on page views, unique visits and returning visits in different browers like chrome , Firefox , opera and many other similar browsers.

III. IMPLEMENTATION OF REGRESSION MODEL

For this study regression model was built in 3 phases [10] -[11] These phases are described below

A. Phase I - Define and Design

In the first phase all study variables were made as clear as possible so that smoother results could be obtained in later phases. It consists of following five steps –

1. Aim of the research questions was written in theoretical and operational terms.

It was decided to study user's behavior depending on the speed of a website.

2. Design the study or define the design

Study was designed to use simple regression model for predicting the bounce rate of webpage. There are many factors that lead to bounce rate of a website some of them may be low quality content, poor website design, slow speed of website may are major factors. It was decided to predict the bounce rate due to slow speed of website as other two factors can't be measured. A simple random sample will be taken from Google webmaster tool.

3. Choose the variables for answering research questions and determine their level of measurement.

Response time of web page was further broken down in to seven variables - Average Page Load Time, Average Server Response Time, Average Page Download Time, Average Redirection Time, Average Server Connection Time and Average Domain Lookup Time. It was decided to get this data Google Webmaster tool.

4. Write an analysis plan

initially data was studied manually for checking the possibility of outliers. Data was preprocessed before making final analysis. Correlation of every variable with the predictor variable that is Bounce Rate was also studied using.

5. Calculate sample size estimations

as the variables to be considered can have different values at different time of day so it was decided to use 3793 rows of data that was recorded for every hour.

B. Phase 2: Prepare and explore

6. Collect, code, enter, and clean data

Data was collected for 2 months in which average of each hour was taken. Here a unique id was attached to each data row which was removed. In some of the attributes response time was given as zero, and in some of the cases time was given more then 5 seconds for a single variable so those rows were also removed. Bounce rate was given in percentage which was converted in to fractions.

7. Run Univariate and Bivariate Statistics

Univariate and Bivariate Analysis were done and anomalies were removed. [12] Distributional assumptions of Yon X were checked for -

a) Independence
b) Normality
c) Constant Variance

It was checked with residual plot- Q-Q plot of the residuals for normality, and a scatter plot of Residuals on X.

8. Run an initial model

Initially model was run to see if any discrepancies are there. Model was examined for outliers and noise.

Phase 3: Refine the model

9. Refine predictors and check model fit

For checking predictor variables effect on response variable i.e. bounce rate, correlation between different variables was checked

10. Test assumptions

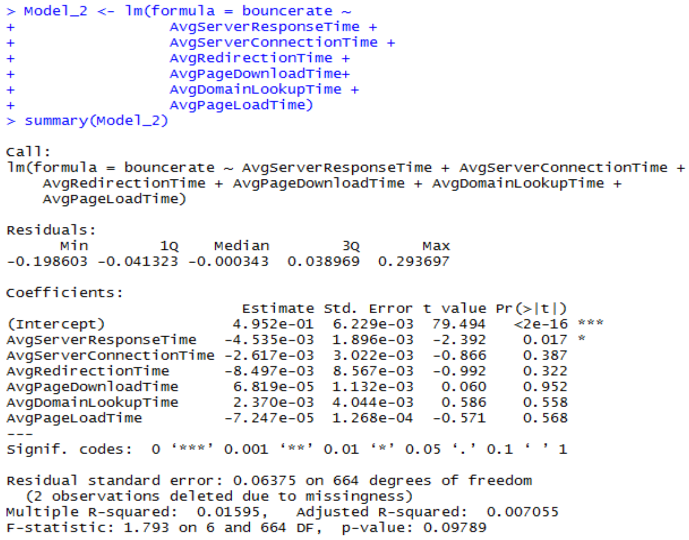
Based on the results obtained in previous step model was built again using 2 variables bounce rate and average Domain Lookup Time.

11. Interpret Results

After model was built and refined mathematical equations were constructed to analyze and predict bounce rate.

IV. RESULTS & DISCUSSION

Initially two months data was collected from Google Analytics which was average of one hour and contained 6742 rows of data.



```
> Model_2 <- lm(formula = bouncerate ~
+                  AvgServerResponseTime +
+                  AvgServerConnectionTime +
+                  AvgRedirectionTime +
+                  AvgPageDownloadTime+
+                  AvgDomainLookupTime +
+                  AvgPageLoadTime)
> summary(Model_2)

Call:
lm(formula = bouncerate ~ AvgServerResponseTime + AvgServerConnectionTime +
    AvgRedirectionTime + AvgPageDownloadTime + AvgDomainLookupTime +
    AvgPageLoadTime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.198603 -0.041323 -0.000343  0.038969  0.293697

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.952e-01  6.229e-03  79.494   <2e-16 ***
AvgServerResponseTime   -4.535e-03  1.896e-03  -2.392    0.017 *
AvgServerConnectionTime -2.617e-03  3.022e-03  -0.866    0.387
AvgRedirectionTime      -8.497e-03  8.567e-03  -0.992    0.322
AvgPageDownloadTime      6.819e-05  1.132e-03   0.060    0.952
AvgDomainLookupTime      2.370e-03  4.044e-03   0.586    0.558
AvgPageLoadTime         -7.247e-05  1.268e-04  -0.571    0.568
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06375 on 664 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.01595,   Adjusted R-squared:  0.007055
F-statistic: 1.793 on 6 and 664 DF,  p-value: 0.09789
```

Fig 1 – Regression Model before Refinement

Retrieval Number: I101030789S19/19©BEIESP
DOI: 10.35940/ijitee.I1103.0789S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This data consisted of attributes like Average Server Response Time, Average Server Connection Time, Average Redirection Time, Average Page Download Time, Average Domain Lookup time and Average Page Load Time. After a close manual examination of data it was discovered that some of attribute values were zero.

Since all the attribute values were measured in seconds and a precision up to 2 decimal places was being used so it was decided to remove rows with zero as value. After preprocessing 4672 rows were left for analysis on which an initial Regression Model was built. Results obtained from this model are depicted in Fig 1

In the result, coefficients are shown in the column Estimate standard dependency between different variables. So, the equation for bounce rate becomes as below.

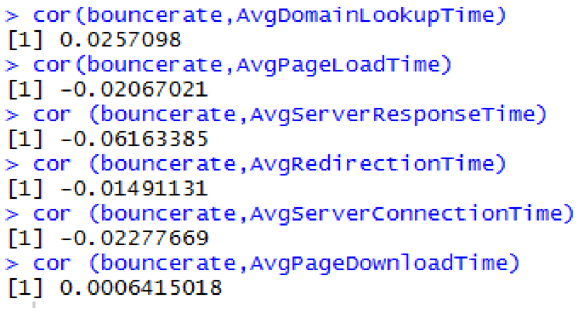
$$bouncerate = 0.5 + (-0.0004)avgServerResponsetime + (-0.0003)\ avgServerconnectionTime + (-0.0008)\ avgRedirectionTime + (0.00007)\ avgPageDownloadTime + (0.002)\ avgDomainLookuptime + (0.00007)\ avgpageLoadtime \qquad (1)$$

Here, from equation (1) it can't be said that the relationships estimated from this regression model(model_2) are perfect, because the model result is generated after model fitted to the data set(i.e. model learns from the data and then estimate coefficients values) and data set may contain some unreliable observations.

It is necessary to improve the model, so that the relationships of bounce rate and time components can be identified very precisely.

It is evident from the equation, Average Domain Lookup Time impacts more on bounce rate. After initial model built various data attributes were checked for correlation among them . Results are depicted in Fig 2.

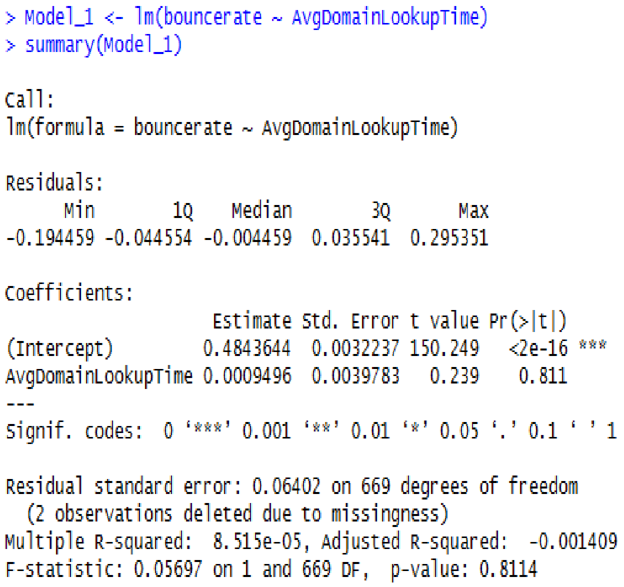
From Fig 2 it was discovered that bounce rate was correlated only with Average Domain Lookup Time, so it was decided to discard other predictor variables.



```
> cor(bouncerate,AvgDomainLookupTime)
[1] 0.0257098
> cor(bouncerate,AvgPageLoadTime)
[1] -0.02067021
> cor (bouncerate,AvgServerResponseTime)
[1] -0.06163385
> cor (bouncerate,AvgRedirectionTime)
[1] -0.01491131
> cor (bouncerate,AvgServerConnectionTime)
[1] -0.02277669
> cor (bouncerate,AvgPageDownloadTime)
[1] 0.0006415018
```

Fig 2 – Checking Correlation between variables

Therefore a new model was built based on bounce rate and Average Domain Lookup Time. Hence the final equation of Bounce Rate is given as –



```
> Model_1 <- lm(bouncerate ~ AvgDomainLookupTime)
> summary(Model_1)

Call:
lm(formula = bouncerate ~ AvgDomainLookupTime)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.194459 -0.044554 -0.004459  0.035541  0.295351 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         0.4843644  0.0032237 150.249   <2e-16 ***
AvgDomainLookupTime 0.0009496  0.0039783   0.239    0.811    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06402 on 669 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  8.515e-05, Adjusted R-squared:  -0.001409 
F-statistic: 0.05697 on 1 and 669 DF,  p-value: 0.8114
```

Fig 3 – Regression Model after Refinement

$$bouncerate = 0.5 + (0.0009)avgDomainLookuptime \qquad (2)$$

It can be seen from equation (2) that Average Domain Lookup Time impacts more on bounce rate.

V. CONCLUSION

For this study it was noticed that there were data issues or misspecified predictor issues until coefficients were interpreted . It was established that to determine the bounce rate of a website its Average Domain Lookup time plays an important role.

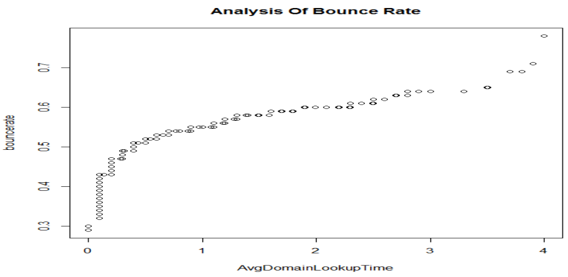


Fig4 qqplot of data after preprocessing

REFERENCES

1. H Singal, S. Kohli, “Trust Necessitated through Metrics: Estimating the Trustworthiness of Websites” Elsevier Procedia Computer Science Volume 85, 2016, Pages 133-140.
2. Dr Veena Dhari, Meenakshi Garg “ Analytics For Content Based Sites –A Comparative Study “ International Journal of Science, Engineering and Technology Research (IJSETR) Volume 5, Issue 6, June 2016 Pages 2261-2264
3. Daniel Amo Filvà, María José Casany Guerrero , Marc Alier

Retrieval Number: I11030789S19/19©BEIESP
DOI: 10.35940/ijitee.I1103.0789S19

Forment, “Google analytics for time behavior measurement in Moodle” 2014 9th Iberian Conference on Information Systems and Technologies (CISTI)

4. laza, Beatriz. “Google Analytics for measuring website performance. Tourism Management - TOURISM MANAGE” 32. 477-481. 10.1016/j.tourman.2010.03.015.
5. Han Qin, Kit Riehle, Haozhen Zhao ” Using Google Analytics to Support Cybersecurity Forensics” 2017 IEEE International Conference on Big Data (Big Data), Dec. 2017,pp: 11-14.
6. https://www.searchenginejournal.com/10-reasons-website-can-high-bounce-rate/182260/#close accessed on 15 May 2019
7. H Singal, S. Kohli, “ Conceptual Model For Obfuscated TRUST induced from Web Analytics data fro content driven Websites” International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014.
8. Akiyuki Sekiguchia, KazuhikoTsuda “ Study on Web Analytics Utilizing Segmentation Knowledge in Business to Business Manufacturer Site “ Elsevier Procedia Computer Science Volume 35, 2014, pp: 902-909
9. Veena Dhari , Meenakshi Garg “A Novel Appraoch for Comparing Third Party Web Analytical Tools for General Data Protection Regulation Policy“ International Journal of Computer Applications 181(42), February 2019, pp: :10-12.
10. Berger P.D., Maurer R.E., Celli G.B. (2018) “ Introduction to Simple Regression. “ In: Experimental Design. Springer, Cham
11. Naomi Altman. Martin Krzywinski “Simple linear regression“, Nature Methods volume12, 2015, pp:999–1000
12. Oyerinde, O. D. , Chia, P. A. , “Predicting Students’ Academic Performances – A Learning Analytics Approach using Multiple Linear Regression“, Vol. 157;No. 4; pp 37- 44

AUTHORS PROFILE

Meenakshi Mttal completed her Bachelors degree in Computer Science in the year 1998 from Kurukshetra University and Masters degree in Computer Science in 2002 from Guru Jambheshwar University in 2002. She completd her Masters degree in Computer Application from Maharishi Dayanand University in 2010. Currently she is pursuing her doctorate degree from RabindraNath Tagore University, Bhopal. Over 15 years of academic experience she has published and presented 15 research papers in various National and International Journals and conferences. Her research interests includes Machine leaning, Data Mining and Software Testing, She has worked as Expert member in various committees in different institutions. She is a member of different professional societies IST, Indian National Congress.

Dr.S.Veenadhari completed her Doctoral programme from Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya in 2015, Master of Computer Applications from Nagarjuna University, Andhra Pradesh in 1998 and Master of Technology in Computer Science and Engineering from Makhan Lal ChaturvediUniversity, Bhopal in 2007. Over 15 years of academic and research experience with 45 research papers published in International and National reputed journals. Six students are pursuing their Doctoral programme and two students completed their doctoral degree under her supervision. Her research interests includes Machine leaning, Big data analytics and Cloud computing. Her expertise in agricultural data analytics through machine learning. She has published several book chapters, technical bulletins, project reports, working papers ,training and teaching reference manuals. She has delivered number invited talks in many national platforms of repute. She has worked as Expert member in various committees in different institutions and member of different professional societies like IE,CSI,ISTE.