CLICKSTREAM PROJECT

Capstone Project Milestone Report - Intro

Clickstream data related to the path the user takes when navigating through the site. During a user journey, a lot of info, articles, and product details will be available to help a user in making a decision. What are the most viewed pages before users take some actions on the page? Is there any relationship between the pages users view and actions they take on the website? Is there any relationship between the pages users view and products they choose on the website?

Goal and Questions

The website is available 24/7 and there are 5 main products available to users. There are also other pages such as articles, product details and services available to help a user choose a product. We are trying to answer some questions such as:

1- Is there any difference between the user behavior in 1st visit versus later visits?

2- Does a number of visit impact on the behaviour? For example, a user with 100 visits vs a user with 20 visits

3- Do we have a path that gives us 100% click rate on the products?

4- Can we profile the user based on the behavior, metrics or dimensions?

5- Is the type of product important?

6- What are the characteristics of users who did not click?

7- What are the characteristics of users who viewed the “result”page and click on the product vs users who did not? Clicks and pageviewes before result matter?

Structure of the Project

The project is structured in 4 big analysis parts.

First, we wrangled the data that we were going to use by cleaning and converting the variables in the dataset. The source of the data is clickstream data for 3 month from Apr 2016 to Jul 2016. We started with 41 variables and 1,031,246 rows. By doing feature engineering, we could have 86 variables and the number of rows went down to 140,336 after cleaning the data.

Second, we explored the data we work on and perform a detailed descriptive analysis}that splits in two different parts. In the first one, the reader can understand what each variable is adding to the dataset and how it is structured. In the second one, we see how each variable is related to our dependent variable, which is “clicking on the product” that helps the reader figure out its relation and helps us understand the importance of each variable for our following analysis. We prepared our final dataset based on this exploratory data analysis and we started the next step with 27 variables and 4555 rows.

Third, the reader can see market basket analysis (association rules)that help us to find some business recommendations because we can see the influence of each variable towards the dependent variable. We fitted our model with 19 variables as a result of previous steps

Fourth, we performed a logistic regression using a binary target variable product_clicked

We would like to mention that, a priori, we expect getting similar conclusions in our different analysis. We believe that since we want to find out which variables are the most important ones in predicting , all the analysis that we are performing tends to give us similar results.

Steps - Review/Explore the dataset

- **We started by answering these questions:**

How the dataset looks? What is the structure? What are variables' formats

- All variable names to lowercase
- **date** variable class is converted from character to date
- **full visitorid** is converted to a factor variable

- **We cleaned the dataset by taking these steps:**

- Initial Visualization: Histograms, plots to see outliers or missing data
- Removed irrelevant columns: **event_value, hits_hit_number, Index, device_is_mobile, total_hits, client_id, plan_id, person_id, session_id, businessid, ssopurpose, hits_is_entrance, hits_is_exit, hits_referer, host, geo_network_country, timestamp, page title, page path, event label, category , action, country, region, hits_page_hostname, hit_timestamp**
- Removed **hit_number** after removing duplicate rows
- Removed rows: undefined events are removed
- Sorted the observations based on viditor_id, visit_number , date and hit_number
- Replace all NA values in **total time** with 0.1
- Filled empty slots with a dummy value 'NA' for **campaign, keyword, ad_content, referral_path, hits_page_page_title,**
- Removed duplicates
- Replaced the values on the "pages" column with the product clicks as "hit" for all the 5 products
- Removed the full visitor ids with no experience on "Gudiance" (/open-account)
- Removed all the rows occurred after the last hit on the result page as we are not interested in knowing what happened after clicking on the product/ after viewing the result page
-

- **Feature Engineering:** We created new variables in the dataset by taking these steps and removed some the insignificant variables when we fitted our model:

- Creating new variables: Based on the dataset, we defined new variables
- New **hit_number** column is added after removing rows

Steps for creating a column for clicking on the product

- New **brokerage_clicked2** column is added if a user has Product-Brokerage-Start-Button event

- New **brokerage_roth_clicked2** column is added if a user has Product-Brokerage-Roth-Start-Button event
- New **emf_clicked2** column is added if a user has Product-EMF-Start-Button event
- New **emf_roth_clicked2** column is added if a user has Product-EMF-Roth-Start-Button event
- New **nonqual_clicked2** column is added if a user has Product-Nonqual-Start-Button event

Creating numeric product variables

- New **brokerage_num_clicked2 column** is added if a user has Product-Brokerage-Start-Button event
- New **brokerage_roth_num_clicked2 column** is added if a user has Product-Brokerage-Roth-Start-Button event
- New **emf_num_clicked2 column** is added if a user has Product-EMF-Start-Button event
- New **emf_roth_num_clicked2 column** is added if a user has Product-EMF-Roth-Start-Button event
- New **nonqual_num_clicked2 column** is added if a user has Product-Nonqual-Start-Button event

Creating number of products clicked variable

- New **product_number_clicked** is added based on all numeric product variables

Creating the 1 target variable

- New **product_clicked** is added based on **product_number_clicked ->**TARGET VARIABLE

- New **app_submitted** column is added if a user submitted an application
- New **had_advisor** column is added if a user has /openaccount/Create-Profile-Advisor in the visit
- New **sso_successfull** column is added if a user logins successfully which means having a page as “/openaccount/Address” or eventlavel as “Advisor-Verification-Code-Success-Continue” when an advisor is helping or if a user comes from “/myVoya/index”
- New **read_article_clean** column is added by grouping all the pages that have articles in url
- New **product_viewed_clean** column is added by grouping all the pages that have products in url
- New **tool_viewed_clean** column is added by grouping all the pages that have tool in url
- New **planning_viewed_clean** column is added by grouping all the pages that have “planning” OR “action” in url
- New **contact_viewed_clean** column is added by grouping all the pages that have contact us in url
- New **had_search_clean** column is added by grouping all the pages that have search in url
- New **had__myVoya_clean** column is added by grouping all the pages that have myVoya in url
- New **entrance** column is added by page title and replaced NAs with page title
- New **exit** column is added by page title replaced NAs with page title
- New **device_is_mobile_2** is created to fix the wrong values in previous device_is_mobile column
- New **maxhit_number variable** is created to find the last hit happened on the result page
- New **total_hits_clean** is created to fix the wrong values in previous total_hits and after removing the duplicate rows

- New **total_pageviews_clean** is created to fix the wrong values in previous total_pageviews
- New **total_events** is created to have the number of events per user and visit
- New **total_resultpageview** is created to find out how many views users had on the result page and eventually how many loops they had
- New **pages** variable is created that has the first 5 characters of each url to work with strings based on "page path" column
- New **article_before_product** is created if a user viewed an article before hitting on the product
- New **article_after_before_prod** is created if a user viewed an article both before and after hitting on the product (Will be not created as the impact was not significant)
- New **product_before_product** is created if a user viewed a product page before hitting on the product
- New **product_after_product** is created if a user viewed a product page after hitting on the product (Will be not created as the impact was not significant)
- New **product_after_before_prod** is created if a user viewed a product page both before and after hitting on the product
- New **planning_before_product** is created if a user viewed a planning page before hitting on the product
- New **tool_before_product** is created if a user viewed a tool page before hitting on the product
- New **contact_before_product** is created if a user viewed a contact-us page before hitting on the product
- New **search_before_product** is created if a user searched before hitting on the product
- New **term_before_product** is created if a user viewed terms before hitting on the product
- New **privacy_before_product** is created if a user viewed privacy before hitting on the product
- New **viewed_term_clean** is created if a user viewed "terms" page(Will be not created as the impact was not significant)
- New **viewed_privacy_clean** is created if a user viewed "privacy" page(Will be not created as the impact was not significant)
- New **viewed_fees_clean** is created if a user viewed fees on product tile(Will be not created as the impact was not significant)
- New **clicked_overview_before** is created if a user clicked on the "overview" as an event label before clicking on the product
- New **clicked_keyfeatures_before** is created if a user clicked on the "keyfeatures" as an event label before clicking on the product
- New **fees_before_product** is created if a user viewed the fees "before" hitting on the product
- New **fees_after_product** is created if a user viewed the fees after hitting on the product (Will be not created as the impact was not significant)
- New **fees_after_before_product** is created if a user viewed the fees before and after hitting on the product(Will be not created as the impact was not significant)
- New **total_articles_before** is created to count the number of articles viewed per user and visit
- New **total_productpageview_before** is created to count the number of product pages viewed per user and visit
- New **total_tools_before** is created to count the number of tools viewed per user and visit

- New **total_planning_before** is created to count the number of planning viewed per user and visit
- New **total_contactus_before** is created to count the number of contact-us viewed per user and visit
- New **total_search_before** is created to count the number of searches per user and visit
- New **total_search_before** is created to count the number of searches per user and visit
- New **total_fee_clicked_before** is created to count the number of privacy viewed per user and visit
- New **total_terms_before** is created to count the number of terms viewed per user and visit
- New **total_hits_onprod** is created to count the number of hits on product per user and visit
- New **diff** variable is created to show how much time each user spent on each page/action
- New **total_time_clean** is created to show how much time users spend in their visit
- New **iraproductoptions_before** is created if a user clicked on **iraproductoptions before** hitting on the product
- New **considerarolloverira_before** is created if a user clicked on **considerarolloverira** before hitting on the product
- New **IRA_before** is created if a user clicked on **IRA before** hitting on the product
- New **PRODUCTS_plusminus_before** is created if a user clicked on **PRODUCTS+- before** hitting on the product
- New **nav_scroll_before** is created if a user clicked on **section-nav-scroll before** hitting on the product
- New **had_vdao_clean** is created to check that all users had vdao in their visit
- New categorical variable **time_on_site** is created based on minutues and statistics(quartiles standard deviation) as ('lessthantwo', 'betweenttwofive', 'betweenfiveten', 'morethanten')
- New categorical variable **totalhit_catg** is created based on total hits and statistics(quartiles standard deviation) as ('(,9]', '[10,20]', '[21,40]', '[41,70]', '[71,]')
- New categorical variable **totalpageview_catg** is created based on total pageviews and statistics(quartiles standard deviation) as ('[1-5]','[6-12]','[13,]')
- New categorical variable **totalevents_catg** is created based on total events and statistics(quartiles standard deviation) as ('[1-7]','[8-19]','[20,]')
- New categorical variable **uniqueproduct_clicked** is created based on the number of unique products that users clicked
- New categorical variable **totalarticle_catg** is created based on total articles users viewed and statistics(quartiles standard deviation) as ('zero','one','[2,]')
- New categorical variable **totalproductpage_catg** is created based on total product page users viewed and statistics(quartiles standard deviation) as ('zero','one','[2,4]','[5,]')
- New categorical variable **totaltoolspage_catg** is created based on total tools users viewed and statistics(quartiles standard deviation) as ('zero','one','[2,]')
- New categorical variable **totalplanning_catg** is created based on total planning pages users viewed and statistics(quartiles standard deviation) as ('zero','one','[2,]')
- New **iraproductoptions_before** created if a user clicked on the “iraproductoptions” as an event label before clicking on the product

Exploratory Data Analysis

- **We did the data visualization using ggplot2 and plotly by taking these steps:**

1- Removed redundant or unnecessary columns :
hit_timestamp, hits_type, hits_page_page_title, hits_page_page_path, event_category, event_action,eventlabel,totals_time_on_site,geo_network_region,referral_path,campaign,keyword ,ad_content,brokerage_clicked2,brokerage_roth_clicked2,emf_clicked2,emf_roth_clicked2,nonqual_clicked2,app_submitted,sso_successfull,read_article_clean,product_viewed_clean,tool_viewed_clean, planning_viewed_clean, contact_viewed_clean,had_search_clean,had_myVoya_clean,exit, device_is_mobile_2,maxhit_number,pages, diff, had_vdao, had_vdao_clean
2- Converted the long format to wide format dataset
3- Remove the observations that spent less than 1 sec on the website
4- -totalplanning_catg, -totaltoolspage_catg are not significant and we don't need them in arules

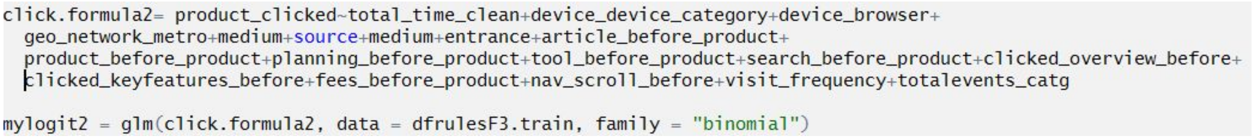
Association Rules (Market Basket Analysis)

1- We dropped the columns that we don't need and prepare the dataset for running "arules" package.
date,total_pageviews_clean,
total_events,brokerage_num_clicked2,brokerage_roth_num_clicked2,emf_num_clicked2,
emf_roth_num_clicked2,nonqual_num_clicked2,product_number_clicked,total_articles_before,
totals_hits_clean,total_resultpageview,total_planning_before,total_contactus_before,
total_search_before,total_productpageview_before,total_tools_before,total_fee_clicked_before,
total_privacy_before,total_terms_before,total_hits_onprod,totalarticle_catg,totalproductpage_catg
totaltoolspage_catg,totalplanning_catg,totalproductpage_catg,totaltoolspage_catg
2- The left hand side columns are:
'device_device_category','device_browser','device_operating_system','device_screen_resolution',
'geo_network_metro','source','medium','had_advisor','entrance',
'article_before_product','product_before_product',
'planning_before_product','tool_before_product','search_before_product','contact_before_product',
'term_before_product','privacy_before_product','clicked_overview_before',
'clicked_keyfeatures_before','fees_before_product','iraproductoptions_before','IRA_before',
'considerarolloverira_before','PRODUCTS_plusminus_before','nav_scroll_before','visit_frequency'
'time_on_site','totalhit_catg','totalpageview_catg'
3-The right hand side will be product_clicked=TRUE& product_clicked=FALSE
4- We ran a shiny app to explore the association rules and do some rules mining

Logistic Regression

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. So, we want to know which variables can help us to predict about our target variable product_clicked.

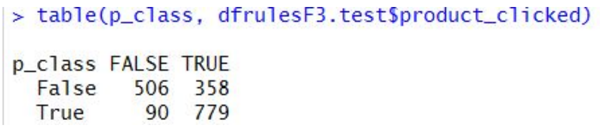
The final variables to fit into our model are:



```
click.formula2= product_clicked~total_time_clean+device_device_category+device_browser+
  geo_network_metro+medium+source+medium+entrance+article_before_product+
  product_before_product+planning_before_product+tool_before_product+search_before_product+clicked_overview_before+
  clicked_keyfeatures_before+fees_before_product+nav_scroll_before+visit_frequency+totalevents_catg

mylogit2 = glm(click.formula2, data = dfrulesF3.train, family = "binomial")
```

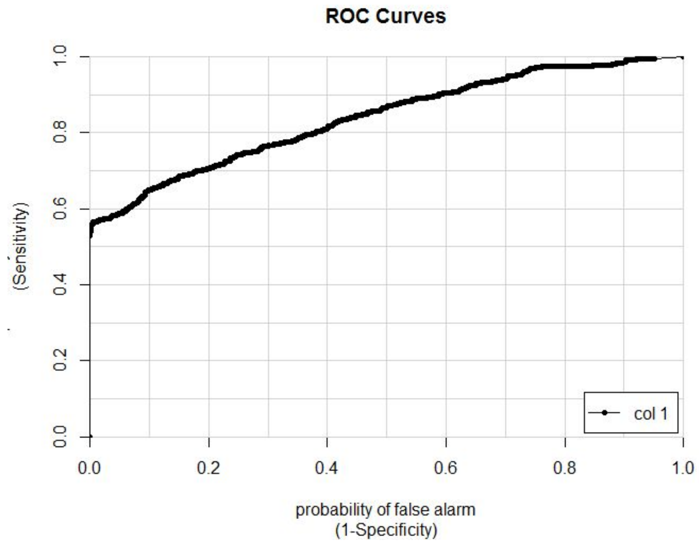
The confusion matrix of the model is:



```
> table(p_class, dfrulesF3.test$product_clicked)

p_class FALSE TRUE
  False   506  358
  True     90  779
```

Our model has the accuracy of 79.5% on the training dataset and the evaluation of the test data shows the accuracy of %75. The sensitivity of our model is 89% and the threshold we chose based on the ROC was 0.65. The AUC was also 0.839 which is a good AUC.



(Because of confidentiality and privacy policy we can’t share more findings on this point)

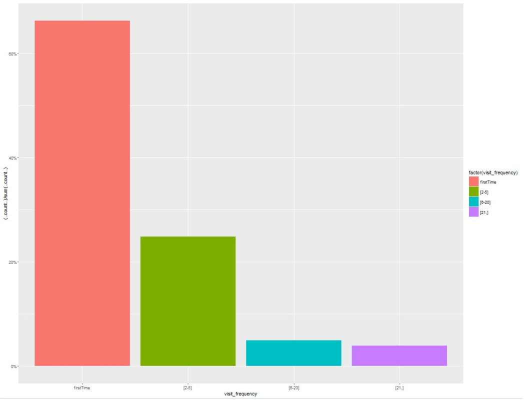
including the model itslef)

Answers to our Questions

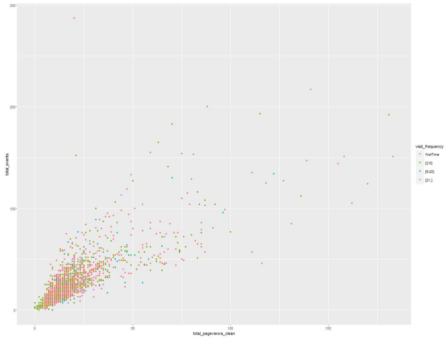
We could find answers to some of our questions.

1- Is there any difference between the user behavior in the 1st visit versus later visits?

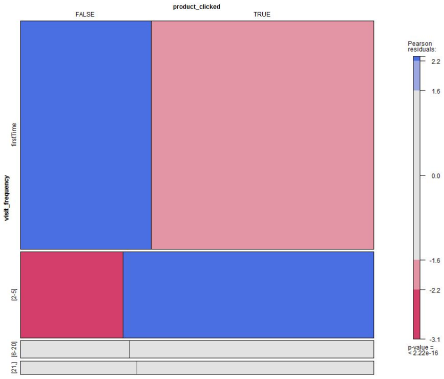
- Our exploratory data analysis shows that visitors who are visiting the website for the first time account for 65% of total visitors.



- They tend to spend more time, view more pages, take more actions.
- We noticed that there is positive correlations between the pageviews and taking actions. The more users view pages the more they take actions.



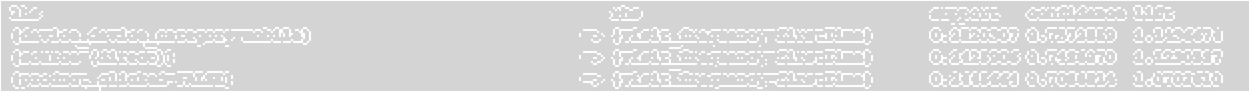
- Mosaic plot shows that the number of first time users who clicked on the products are significantly higher than the first timers who did not click on the product.



- Market Basket Analysis: some of most important rules are:

- 75% of mobile users visited the website for the first time which was 14% more than what was expected

- 70% of no clicks on the products were done by the first timers

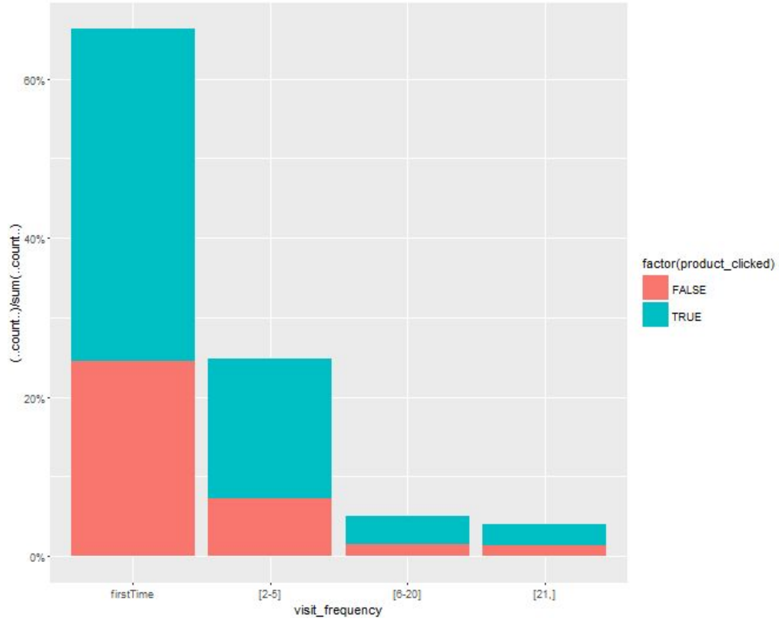


```
lhs                                      rhs                            support   confidence lift
{device_device_category=mobile}       => {visit_frequency=firstTime}   0.2120307 0.7572860  1.1424671
{source=(direct)}                     => {visit_frequency=firstTime}   0.2425003 0.7431970  1.1220167
{product_clicked=FALSE}               => {visit_frequency=firstTime}   0.2116644 0.7031424  1.0702130
```

- Logistic regression: Our model shows that being a first timer on the website does not predict the click on the product and this variable is not statistically predictive variable. This also confirms our findings in market basket analysis.

2- **Does a number of visit impact on the behavior? For example, a user with 100 visits vs a user with 20 visits?**

- By grouping the visit frequency number to different categories, we realized that users with medium/high frequency of visit, spend less time on the site, view fewer pages and take fewer actions that first-time users.



- Our logistic regression model shows there is a significant difference between the behavior of users with the high frequency of visit first-time users and low/medium visit frequency users. Medium frequency users have negative coefficient and this might have different reasons. Some of them might have already clicked on the product in their previous visits.

3- Do we have a path that gives us 100% click rate on the products?

Both market basket analysis and exploratory data analysis shows that there are a specific group of users that took specific paths and got 100% conversion rates. Our regression model also shows their characteristics are unique and statistically significant. For example, some of them started their journey from a specific path and some of them referred by the specific sources. (Because of confidentiality and privacy policy we can’t share more findings on this point)

4- Can we profile the user based on the behavior, metrics or dimensions?

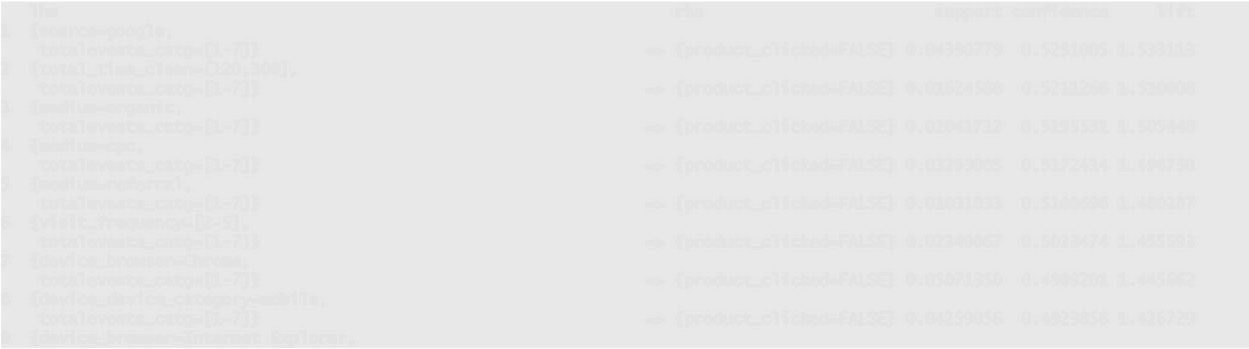
Both market basket analysis and exploratory data analysis shows that there are a specific group of users that took specific paths and got 100% conversion rates and also there are some unique characteristics such as source and clicking on some features on the page that impact the user behavior in terms of not clicking on the product.

5- Is the type of product important?

In order to answer this question, we will need further analysis specific to the type of products. This can be part of our future analysis and recommendations.

6- What are the characteristics of users who did not click?

Our market basket analysis shows that there are some dimensions such as source, device browser, visit frequency and device category that impact the clicks on the product. (Because of confidentiality and privacy policy we can't share more findings on this point)



```
   lhs                                                              rhs                     support    confidence lift
1  {source=google,
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.04390779 0.5291005 1.533103
2  {total_time_class=[120,300],
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.01624598 0.5211268 1.510008
3  {medium=organic,
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.02041732 0.5195531 1.505448
4  {medium=cpc,
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.03293085 0.5172414 1.498750
5  {medium=referral,
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.03031833 0.5108696 1.480287
6  {visit_frequency=[2-5],
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.02349067 0.5023474 1.455592
7  {device_browser=Chrome,
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.05071350 0.4989201 1.445662
8  {device_device_category=mobile,
    totalevents_catg=[1-7]}                                      => {product_clicked=FALSE} 0.04259056 0.4923858 1.426729
9  {device_browser=Internet Explorer,
```

7- What are the characteristics of users who viewed the "result"page and click on the product vs users who did not? Do clicks and pageviewes before result matter?

We already answered to some part of this question in the previous questions. Viewing some section of the website such as articles and product detail pages impact the later behavior of the users in terms of clicking. One of the key findings was if we direct some of the incoming traffic to specific pages and users start their journey from those pages, we will have better conversion rates and higher clicks on the product. Based on our logistic regression, assuming all the coefficients are equal, if we increase the number of visits to 3 specific pages by 1, we will increase the number of clicks to 4 for each entrance.

Recommendations

1- For paid traffic and paid ads, it is better to link the traffic to specific product pages as they will increase the engagement and product familiarity which will lead to more clicks on the products.
2- There are specific blogs, referrers, and websites that send visitors to the pages. There is a good opportunity to invest money in this segment as they have significant impact on the conversions
3- There is no need to change the flow in other sections such as articles, tools and contact us as they have no statistically significant impact on the conversions.
4- Traffic that comes from the emails have high statistically significant conversions and further investment may be a good choice on targeting the customers via email.
5- New visitors spend more time on the pages and there is an opportunity for the company to increase their conversions by showing them the product detail pages and educate them. This will lead to high-quality conversions and clicks on the products.
(Because of confidentiality and privacy policy we can’t share more findings on this point)

Further Research

There should be a further study on the visitors who visit the website more than 6 times but less than 20. This segment has no interests on the product itself but the fact that they are visiting the pages, again and again, shows they are looking for other info. Understanding these visitors may have some good impacts on their user experience.
Any analysis to measure ROI, any marketing activities such as online ads, social media posts and marketing automation process such as weekly, daily and monthly emails should be targeted based on web analytics modeling to increase the conversions and to measure the effectiveness of marketing channels.

Appendix

Variables (columns) are:

hit_timestamp	STRING
date	STRING
fullVisitorId	STRING
visitNumber	INTEGER
hits_type	STRING
hits_hitNumber	INTEGER
totals_pageviews	INTEGER
totals_hits	INTEGER
hits_page_pageTitle	STRING
hits_page_pagePath	STRING
EventCategory	STRING
EventAction	STRING
Eventlabel	STRING
totals_timeOnSite	INTEGER
ClientID	STRING
PlanID	STRING
PersonID	STRING
SessionID	STRING

SSOPURPOSE	STRING
EventValue	INTEGER
hits_page_hostname	STRING
device_deviceCategory	STRING
device_browser	STRING
device_operatingSystem	STRING
device_isMobile	BOOLEAN
device_screenResolution	STRING
geoNetwork_country	STRING
geoNetwork_region	STRING
geoNetwork_metro	STRING
hits_isEntrance	BOOLEAN
hits_isExit	BOOLEAN
hits_referer	STRING
Host	STRING
ReferralPath	STRING
Campaign	STRING
Source	STRING
Medium	STRING
Keyword	STRING
AdContent	STRING