Internet Sports Gambling Activity Marketing Analysis

Wajih ARFAOUI & Vinay RAJAGOPALAN

2021-12-21

Contents

Project's objective

The aim from this project is to provide the marketing team with insights about the users of the company's gambling websites, using datasets collected from the collaborative Internet gambling research project between the **Division on Addictions (DOA)** and **bwin Interactive Entertainment, AG (bwin)**.

Data Manipulation

In order to come with a **marketing data mart** through which we can capture the customer's behaviors and an overall picture of his profile, we used these datasets:

- **Demographics**: contains demographic information per `User ID` such as: `Gender`, `Language`, and `Registration Date`.
- **UserDailyAggregation**: contains the actual betting information associated with each product for each participant for each calendar day. For instance we have: `Betting Product`, `Stakes`, `Winnings`, and `Bets`.
- **PokerChipConversions**: contains the actual poker chip transaction information from February 1, 2005 through September 30, 2005. The poker transaction information includes `User ID`, `Transaction Date/Time`, `Transaction Type (buy or sell)`, and `Transaction Amount`.

After loading these datasets, we moved to **data preparation step**, in which we made sure that these tables are clear and ready to be used, through checking *missing values*, *outliers* and that our *variables have the correct data types*.

Once the three tables are ready, we applied a function on the **UserDailyAggregation** dataset to group the betting information by `User ID` to get a table with only one row of information per unique customer.
During this step, we thought about applying the ***RFM Analysis technique*** to create new variables for each `User ID` such as `Active_Days` (Recency), `sum_Stakes` (Monetary), and `count` (Frequency) on an overall and product's levels.
Here is an overview of the basetable we got after applying the aggregation functions and merging the datasets together:

Table 1: Final DataMart

UserID	RegDate	FirstPay	FirstAct	FirstSp	sum_Stakes	count	Last_date_played	First_date_played	Active_Days
1324354	2005-02-01	2005-02-24	2005-02-24	2005-02-24	11976.6100	136	2005-09-30	2005-02-24	218
1324355	2005-02-01	2005-02-01	2005-02-01	2005-02-01	425.5600	106	2005-09-29	2005-02-01	240
1324356	2005-02-01	2005-02-01	2005-02-02	2005-02-02	1365.2600	75	2005-09-12	2005-02-02	222
1324358	2005-02-01	2005-02-01	2005-02-01	2005-02-01	336.2898	9	2005-05-06	2005-02-01	94
1324360	2005-02-01	2005-02-02	2005-02-02	2005-02-02	65.7427	32	2005-09-25	2005-02-02	235
1324362	2005-02-01	2005-02-11	2005-02-11	2005-02-11	22.0000	7	2005-09-17	2005-02-11	218
1324363	2005-02-01	2005-02-01	2005-02-01	2005-02-01	275.5300	6	2005-02-22	2005-02-01	21
1324364	2005-02-01	2005-02-03	2005-02-03	2005-02-03	295.0000	22	2005-09-28	2005-02-03	237

```
## [1] 42648   171
```

After this, we moved to do **segmentation** to assign each customer to a specific cluster based on score we calculated using the RFM values:

$RFM\_score = 100 * R\_score + 10 * F\_score + M\_score$

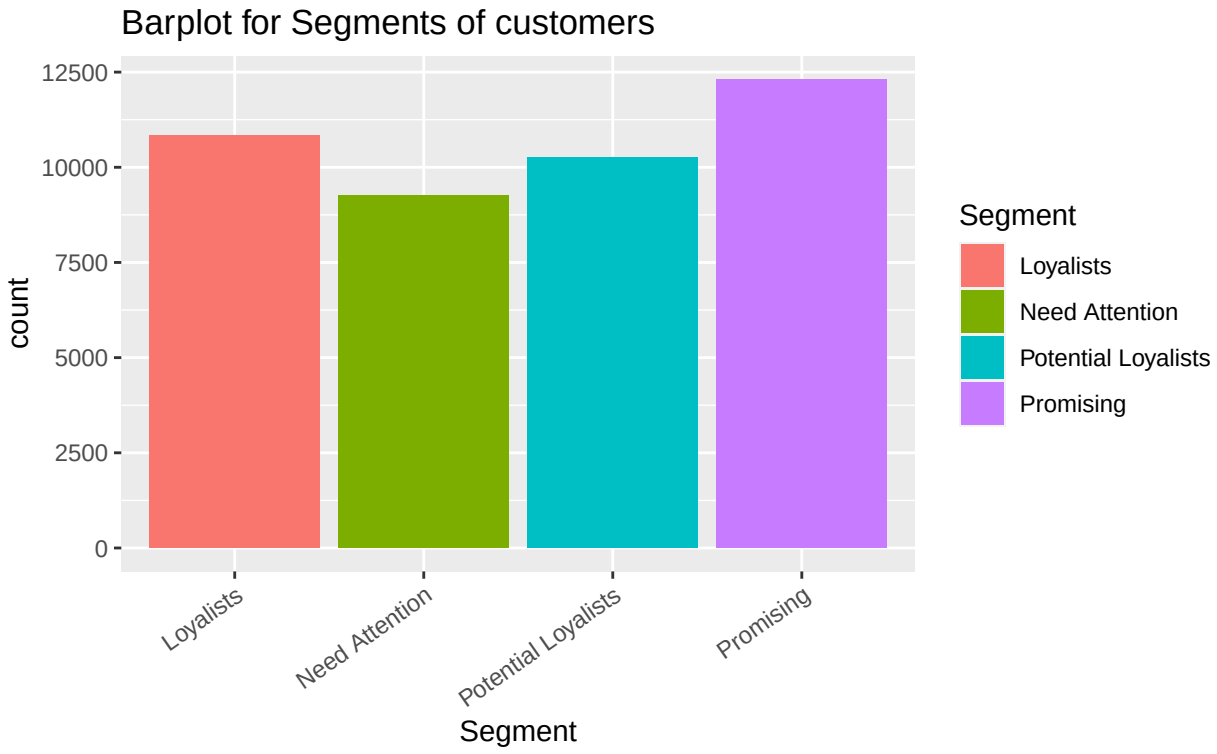
where:
- **R_score** is the Recency Score
- **F_score** is the Frequency Score
- **M_score** is the Monetary Score

And we ended up by having RFM scores distributed like this:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   143.0   243.0   270.5   343.0   430.0
```

Based on these values, we decided to create four clusters:
- **Loyalists**: with `RFM_Score` between 430.0 & 343.0
- **Potential Loyalists**: with `RFM_Score` between 343.0 & 270.5
- **Promising**: with `RFM_Score` between 270.5 & 143.0
- **Need Attention**: with `RFM_Score` less than 143.0

And we had our customers' distribution like this:



After we succeeded to have an overview of our customers' segments, we thought about digging deeper by doing **kmeans** clustering for each product inside the overall clusters.
Each product's customers were split into four clusters:
- **High Value** customers - **Medium Value** customers - **Low Value** customers - **Never Played** customers

Finally, we exported the basetable in CSV format to be used after for the shiny app.

Shiny App

Please click here to access our Shiny App

In order to make it more user friendly to explore the data of this basetable, we created the shiny interactive app that will allow you to apply filters, choose the variables you want to visualize and get a snapshot of the data mart with the possibility of downloading it under different formats.

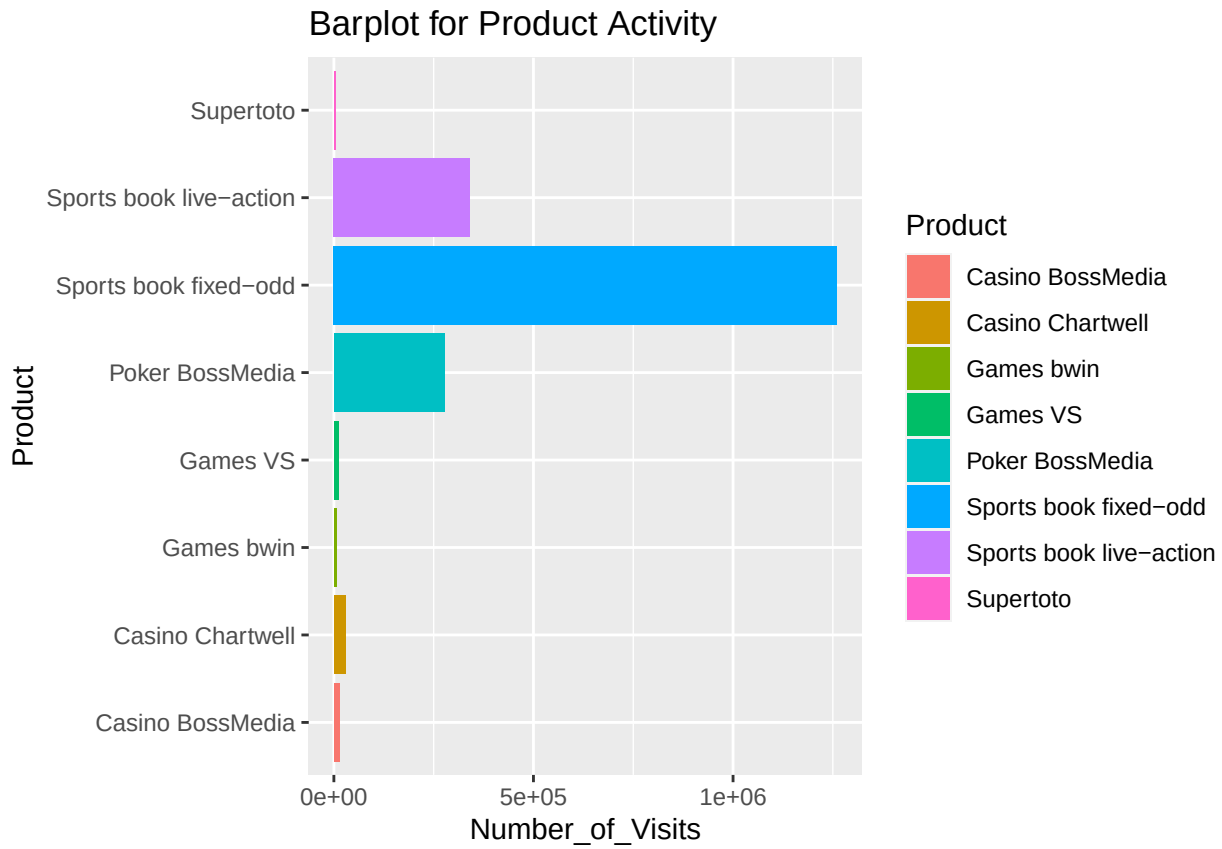
The dashboard structure

Tabs: There are 5 different tabs that allow the user to explore different insights from the data described below:

- **Demographical Insights**: contains graphs showing the distribution of customers per categorical variables such as `Language`, `ApplicationID`, and `Gender`.
- **Behavioral Insights**: contains graphs highlighting the share of different monetary variables over the categorical ones. For instance, we have the heat map showing the sum_stakes per country.
- **Overall Insights**: In this tab, we plot categorical variables against each other with a numeric variables as size to come up with a bubble plot. In addition to that, we included a summary table to show the different statistic values of the size variable.
- **Product's Insights**: this tab shows a bar chart for the 4 different clusters of customers based on overall RFM score by default. We have also the possibility to select a specific product to get the subclusters distribution of this product inside the overall clusters.
- **Data Mart**: contains a table that contains the data for the selected product. This data can be filtered by the different variables and also can be extracted under different formats as mentioned before.

Insights

- The majority of **bwin** customers are Males from Germany who are using mostly *betandwin.de* and *betandwin.com* to gamble.
- The mean net profit for all the products is `-190.49 euros` with Germany being the country with the highest total amount of losses.
- The mean active days for customers across all the products is `39 days`.
- The customers with the highest average winnings are **Italian males** with a value of `9,076 euros`, followed by **Austrian Females** with an average winnings of `5,718 euros`.
- The top 3 products that customers most bet on are: **Sports book fixed-odd (Prod1)**, **Sports book live-action (Prod2)**, and **Poker BossMedia (Prod3)** respectively as shown in the graph below.



- Based on our segmentation, we can see that most of the existing customers are **promising loyalists** 29% followed by **Loyalists** 25%.

```
##
##          Loyalists      Need Attention Potential Loyalists           Promising
##              10832                9253               10249               12314
```

- The leading products' cluster constructing our overall segments is the **low_value** cluster which groups customers with relatively low RFM *(Recency-Frequency-Total_bets)* values, followed by **medium_value** cluster for customers with RFM vlaues that are around average when compared to the overall values for that product.

References

- https://dreamrs.github.io/shinyWidgets/reference/pickerOptions.html
- https://shiny.rstudio.com/articles/datatables.html
- https://stackoverflow.com/questions/53499066/downloadhandler-with-filtered-data-in-shiny
- http://shinyapps.dreamrs.fr/shinyWidgets/
- https://search.r-project.org/CRAN/refmans/shinyWidgets/html/pickerInput.html
- https://stackoverflow.com/questions/33488924/adding-multiple-conditions-in-conditionalpanel-in-shiny
- https://medium.com/analytics-vidhya/customer-segmentation-using-rfm-analysis-in-r-cd8ba4e6891
- https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/
- https://towardsdatascience.com/k-means-clustering-concepts-and-implementation-in-r-for-data-science-32cae6a3ceba
- https://www.business-science.io/business/2016/09/04/CustomerSegmentationPt2.html/