Open Source Programming Group Assignment: BWIN

Group 9: Patrick Dundon, Ruturaj Mokashi, Priya Varadarajan

December 13, 2018

Introduction

This project is designed to provide marketing insights for the online gambling data collected by the collaborative Internet gambling research project between the Division on Addictions (DOA) and bwin Interactive Entertainment, AG (bwin), an Internet betting service in Vienna, Austria.

The data analyzed is from the period of February 1, 2005 through September 30, 2005. Among our main focuses include aggregated betting behaviour from over 40,000 bwin Internet sports subscribers and players of other online casino games.

The objective of our project would be to provide some useful marketing insights from the datasets provided and create a basetable which can be used for further predictive analytics.

Brain storming

With each raw dataset, we decided which variables to keep, drop, create and finally merge with our final basetable (marketing data mart). Then, once the final basetable was complete, we also created a final table for each of Internet sports betting, poker, casino, games, supertoto and general customer demographics, which were used to generate useful insights for every game in our R Shiny application.

The Basetable creation (Marketing Data Mart)

The basetable contains all the most important variables from all subsections of our data and hence it is the largest table in our project. It was derived initially from merging the user daily aggregation and general demographic raw datasets, which provided us with basic variables such as: specific customer UserID, Gender and Continent* (individual user country was replaced by their respective continent for the purpose of convinient analysis and was implemented using the countrycode package in R):

```
library(countrycode)
Gambdemo1$Continent <- factor(countrycode(sourcevar = Gambdemo1[, "Names"],
                                          origin = "country.name",
                                          destination = "continent"))
```

With the backbone of the basetable now complete, we began to bring in elements of each smaller table to reach our final amount of variables (approx. 35). You will see a lot of variables in further sections referring to maximum, mean and recency of different user activities corresponding to every product. These were intially calculated for the creation of the basetable.

To begin calculating these basetable variables, we had to convert the date format of user activity to proper date format for calculating the variables such as Recency. This was done by using the separate function from the *tidyr* R package and the str_pad function from the *stringr* R package and also date manipluation functions like as.date. This displays date in a readbale format for future use.

Variables such as max, mean was calculated for stakes,bets and winnings and recency were then computed for all products (labelled as products 1-8 and later renamed). In order to calculate recency we subracted the max date of betting activities from 1st October 2005 in order to include the data of betting activities on Sep 30th.:

```
Gambagg2 <- Gambagg1 %>% group_by(UserID,ProductID) %>% summarise(mean_Stakes = round(mean(Stakes),2),
                              mean_Winnings = round(mean(Winnings),2),
                              mean_Bets = round(mean(Bets),2),mean_12Stakes =
                              round(max(Stakes),2),mean_Recency_bet = as.Date("01/10/2005",
```

```
                                    '%d/%m/%Y') - max(Date))

#This was used to spread the variables for every user across products
Gambagg3 = Gambagg2 %>%
  gather(Var, val, starts_with("mean")) %>%
  unite(Var1,Var, ProductID) %>%
  spread(Var1, val)
```

In the Poker data set, originally the data was grouped by userid and Pokersell/Buy value. This was then spread for Poker buy and sell values into separate columns and profit (substracting buy in from cash out amounts) was calculated, which would yield one of important marketing insights:

```
Gambpok1 <- Gambpok %>% group_by(UserID,TransType) %>% summarise(mean_TransAmount =
round(sum(TransAmount),2),mean_Recency_poker = as.Date("01/10/2005",'%d/%m/%Y') - max(TransDateTime))

Gambpok2['Profit_Poker'] = Gambpok2$Poker_Sell - Gambpok2$Poker_Buy
```

After each stage of calculating new variables for each product, we merged that table with the previous table, which in essence was building our basetable block by block; In the same step, any columns that were deemed uneccessary for analysis were later removed. Overall, we merged the aggregation table, Poker and demographics table in order to create the final basetable.

The variable continent was converted to dummy variables (either 0 or 1) across continents so as to be suitable for prediction models and that is how they appear in the basetable.

However, for analysis and graphical representation, these variables were reduced back to one column each with character values “Male/Female” or “Europe/Asia”, etc. Also, other values that displayed as NA were changed to 0 to ensure further analysis would not give any errors. Recency columns which had a negative value was also converted to 0 as they did not have any significance.

Finally, as explained in more detail in the Casino and Games sections, we merged related variables for products 4 and 8 together (both casino) and products 6 and 7 together (both games). An example of how this was done is seen here:

```
Gambdataoveralldata1['Casino_max_stakes'] = Gambdataoveralldata1['Max_Stakes_4'] +
Gambdataoveralldata1['Max_Stakes_8']
Gambdataoveralldata1['Games_max_stakes'] = Gambdataoveralldata1['Max_Stakes_6'] +
Gambdataoveralldata1['Max_Stakes_7']
Gambdataoveralldata1$Max_Stakes_4 = Gambdataoveralldata1$Max_Stakes_8 =
Gambdataoveralldata1$Max_Stakes_6 = Gambdataoveralldata1$Max_Stakes_7 = NULL
```

After renaming columns to accurately reflect what they represent and re_arranging the order so that each product is grouped together, we now had our completed basetable:

```
##   X  UserID Gender sportsFO_max_stakes sportsLA_max_stakes
## 1 1 1324354      1              680.00              392.62
## 2 2 1324355      1               32.53                6.70
## 3 3 1324356      1              150.00              140.80
## 4 4 1324358      1               73.45               88.59
## 5 5 1324360      1               12.85                1.15
## 6 6 1324362      1                5.00                0.00
##   sportsFO_mean_Bets sportsLA_mean_Bets sportsFO_mean_Stakes
## 1               2.02                2.26                86.64
## 2               2.33                3.00                 4.05
## 3               1.92                4.83                13.45
## 4               0.88                4.00                30.96
## 5               1.38                1.50                 2.07
## 6               1.00                0.00                 3.14
```

```
##   sportsLA_mean_Stakes sportsFO_mean_Winnings sportsLA_mean_Winnings
## 1                96.80                  87.39                  79.60
## 2                 3.53                   4.58                   1.60
## 3                28.29                   5.60                  26.05
## 4                88.59                  19.23                  55.98
## 5                 0.87                   1.38                   0.60
## 6                 0.00                   0.00                   0.00
##   Recency_bet_sportsFO Recency_bet_sportsLA Supertoto_max_stakes
## 1                    1                    5                    0
## 2                    2                  236                    0
## 3                   19                   20                    0
## 4                  148                  151                    0
## 5                    6                   10                    0
## 6                   14                    0                    0
##   Supertoto_mean_Bets Supertoto_mean_Stakes Supertoto_mean_Winnings
## 1                   0                     0                       0
## 2                   0                     0                       0
## 3                   0                     0                       0
## 4                   0                     0                       0
## 5                   0                     0                       0
## 6                   0                     0                       0
##   Supertoto_LOS Casino_max_stakes Casino_mean_bets Casino_mean_Stakes
## 1             0                 0                0                  0
## 2             0                 0                0                  0
## 3             0                 0                0                  0
## 4             0                 0                0                  0
## 5             0                 4                4                  4
## 6             0                 0                0                  0
##   Casino_mean_Winnings Casino_LOS Games_max_stakes Games_mean_bets
## 1                    0          0                0               0
## 2                    0          0                0               0
## 3                    0          0                0               0
## 4                    0          0                0               0
## 5                    2        240                0               0
## 6                    0          0                0               0
##   Games_mean_Stakes Games_mean_Winnings Games_LOS Poker_Sell Poker_Buy
## 1                 0                   0         0       0.00      0.00
## 2                 0                   0         0      30.83      8.26
## 3                 0                   0         0       0.00      0.00
## 4                 0                   0         0       0.00      0.00
## 5                 0                   0         0       0.00      0.00
## 6                 0                   0         0       0.00      0.00
##   Profit_Poker Recency_Poker Africa Americas Asia Europe Oceania
## 1         0.00             0      0        0    0      1       0
## 2        22.57           108      0        0    0      1       0
## 3         0.00             0      0        0    0      1       0
## 4         0.00             0      0        0    0      1       0
## 5         0.00             0      0        0    1      0       0
## 6         0.00             0      0        0    0      1       0
##   Others_Country
## 1              0
## 2              0
## 3              0
## 4              0
```

```
## 5              0
## 6              0
```

Now that the basetable was complete, we were able to merge the it back with the other raw data tables (sports betting, poker, casino, Supertoto and other games) to construct smaller specific tables as well. This is detailed in the sections below.

Internet Sports Gambling Table

We wanted to do some extensive analysis on the Sports Gambling table. Hence new variables from the raw sports gambling table were derived for both Fixed-Odds and Live Action in order to tell us more about the specific average betting stake, average winnings and recency of use for each customer. Average stake was calculated by dividing previously existing columns FOTotalBets and LA TotalBets from FOTotalStakes and LATotalStakes respectively.

```
SportsBetting$FOAvgStake <- SportsBetting[,'FOTotalStakes']/SportsBetting[,'FOTotalBets']
SportsBetting$LAAvgStake <- SportsBetting[,'LATotalStakes']/SportsBetting[,'LATotalBets']
```

The Fixed-Odds and Live Action average winnings variables were derived from dividing FA/LATotalBets from FO/LATotalWinnings:

```
SportsBetting$FOAvgTotalWinnings <- SportsBetting[,'FOTotalWinnings']/SportsBetting[,'FOTotalBets']
SportsBetting$LAAvgTotalWinnings <- SportsBetting[,'FOTotalWinnings']/SportsBetting[,'LATotalBets']
```

The Fixed-Odds and Live Action recency variables were calculated by subtractive the usrs last active date from the end date of the data collection (September 30th, 2005):

```
SASDate_Sept302005 <- 16710
SportsBetting$FORecency <- SASDate_Sept302005 - SportsBetting$FOLastActiveDate
SportsBetting$LARecency <- SASDate_Sept302005 - SportsBetting$LALastActiveDate
```

All NaN values were also replaced with 0, so as to be able to manipulate data later on. The gender column was also changed from 0 and 1 to “Female” and “Male” respectively. Columns like language and registration were dropped as they were deemed less relevant than continent and recency. At this point we had the neccessary information to evaluate all users habits relating to spending, success and how often they play which are invaluable to understanding the customer base. The final sports gambling table therefore is the following:

```
##   X  UserID AGE FOAvgStake LAAvgStake FOAvgTotalWinnings
## 1 1 1324354  42  42.954788 42.7739535          43.322542
## 2 2 1324355  22   1.735325  1.1761905           1.962338
## 3 3 1324356  28   7.001939  5.8540517           2.913367
## 4 4 1324358  24  35.385300 22.1481750          21.982229
## 5 5 1324360  27   1.499982  0.5811333           0.998910
## 6 6 1324362  22   3.142857  0.0000000           0.000000
##   LAAvgTotalWinnings FOTotalActiveDays LATotalActiveDays Gender
## 1         35.1741861               218               145   Male
## 2          0.5333333               240                 6   Male
## 3          5.3892241               222               221   Male
## 4         13.9954750                94                 0   Male
## 5          0.4003000               235               221   Male
## 6          0.0000000               218                 0   Male
##   Recency_bet_sportsFO Recency_bet_sportsLA Continent
## 1                    1                    5    Europe
## 2                    2                  236    Europe
## 3                   19                   20    Europe
## 4                  148                  151    Europe
## 5                    6                   10      Asia
## 6                   14                    0    Europe
```

Poker Table

As the poker player data was smaller in nature (approx. 2,000 users) and there were few performance variables, we needed to wait until the final basetable was complete to make the final poker table.

Final variables for this table included UserID, Gender (0 and 1 replaced with "Female" and "Male"), Poker_Sell and Poker_Buy (origanal variables from raw data), Profit_Poker (Poker_sell - Poker_Buy), Recency_Poker and Continent. We found as a smaller dataset, it was best to look at the basics: How much players were spending, whether they were making a profit or loss, how often they play and their location.

The final poker table is shown below:

```
##    UserID Gender Poker_Sell Poker_Buy Profit_Poker Recency_Poker Continent
## 1 1324355   Male      30.83      8.26        22.57           108    Europe
## 2 1324368   Male    9161.66   9620.81      -459.15           124    Europe
## 3 1324369   Male       0.29      0.29         0.00           212      Asia
## 4 1324372   Male     860.78    918.41       -57.63             0    Europe
## 5 1324377   Male      27.55     12.99        14.56           242    Europe
## 6 1324379   Male       0.96      0.88         0.08           164    Europe
```

Casino Table

We decided to merge the two casino products offered and tracked for data into one cluster for analysis. In addition to the common UserID, gender and continent variables, we also calculated specific user metrics such as: max stakes, average bets, average stakes, average winnings and length of subscription. These variables allow us to see the length of platform use, volume of bets as well as whether the user is making a profit or a loss.

```
##    UserID Gender Casino_max_stakes Casino_mean_bets Casino_mean_Stakes
## 1 1324360   Male               4.0             4.00               4.00
## 2 1324363   Male             104.0            46.33              78.00
## 3 1324368   Male             364.5           163.50             231.75
## 4 1324369   Male             618.0           120.60             152.00
## 5 1324372   Male               2.0            12.00               2.00
## 6 1324377   Male              15.0             2.00              15.00
##   Casino_mean_Winnings Casino_LOS Continent
## 1                 2.00        240      Asia
## 2                60.50        224    Europe
## 3               212.75         38    Europe
## 4               148.40        151      Asia
## 5                 2.00         14    Europe
## 6                 4.00        180    Europe
```

Supertoto Table

A separate table and tab was created for Supertoto due to it's specific nature in comparison with the other products offered. Similarly to the Casino table above, Supertoto includes the base UserID, gender and continent measures along side calculated variables for max stakes, average bets, average stakes, average winnings and length of subscription.

```
##    UserID Gender Supertoto_max_stakes Supertoto_mean_Bets
## 1 1324386   Male                  3.2               1.40
## 2 1324454   Male                 11.4               1.60
## 3 1324527   Male                  5.5               2.00
## 4 1324708   Male                  3.3               2.09
## 5 1324866   Male                  1.5               3.50
## 6 1324868   Male                  0.1               1.00
```

```
##   Supertoto_mean_Stakes Supertoto_mean_Winnings Supertoto_LOS Continent
## 1                  0.92                    0.76           50    Europe
## 2                  3.21                    2.65          208    Europe
## 3                  5.50                    4.54           51      Asia
## 4                  0.98                    0.81          115    Europe
## 5                  0.81                    0.68           80    Europe
## 6                  0.10                    0.08          149      Asia
```

Games Table

This table summarizes user statistics on all other games of which data was collected (Games VS, Games bwin). Similar to the other tables, it provides metrics for measuring customer betting habits and amount of time using the platform. The specifc metrics include UserID, gender and continent measures along side calculated variables for max stakes, average bets, average stakes, average winnings and length of subscription.

```
##    UserID Gender Games_max_stakes Games_mean_bets Games_mean_Stakes
## 1 1324368   Male             1.00            1.00              1.00
## 2 1324369   Male             9.85            8.00              9.85
## 3 1324372   Male            16.57            9.50              9.29
## 4 1324379   Male           245.93           35.42             84.14
## 5 1324383   Male            22.10            6.00             22.10
## 6 1324386   Male           248.31           45.33             89.63
##   Games_mean_Winnings Games_LOS Continent
## 1                0.00        41    Europe
## 2                4.90       217      Asia
## 3                9.04        29    Europe
## 4               78.20        32    Europe
## 5                2.04       114    Europe
## 6               74.86         1    Europe
```

Demographics Table

We also deemed it important to create a demographics table to provide a clear picture of the users with regards to gender and location. Once we had determined the continent of each user, the table was made quite easily as gender and UserID were already established. Data from the basetable was extracted in order to do this. Again, gender values were changed from 0 and 1 to Female and Male. See the table below:

```
##    UserID Gender Continent
## 1 1324354   Male    Europe
## 2 1324355   Male    Europe
## 3 1324356   Male    Europe
## 4 1324358   Male    Europe
## 5 1324360   Male      Asia
## 6 1324362   Male    Europe
```

Shiny Application

In order to view and analyse our marketing metrics for the gambling dataset, we have created a Shiny R application with multiple tabs where you can select and plot the different variables to gain an understanding how customers behave in different markets and games. The first tab you will see, entitled **Sports**, will allow you to compare the Fixed-Odds and Live Action performance of each user, with filters available for gender and continent. For example, the interactive plot on this tab allows you to evaluate user consumption through how much they are winning (or not winning) as well as betting habits. There are also a couple of fixed histograms on the tab which show user loyalty by displaying the recency of users for both Fixed-Odds and Live Action platforms.

Next, we created a **Sports(StakesVSAge)** tab to visualize the relationship between the Fixed-Odds and Live Action average stakes and winnings across all the age groups within sports betting. The inference is that people of age groups between 25 to 55 are the ones who have played the most.

For the first two tabs, the user can play around with the left-side panel of the shiny app, adjusting the X and Y axes to compare different variables. Filters include gender (Male and Female) and continents. The alpha factor can be increased or decreased to enhance the pixels. Beyond the first two tabs, the user will not have ability to change the X and Y axes (these will be fixed to certain variables), but the filters for gender and continent will still be applicable. Now, we will explain the rest of the tabs available (each corresponding to a different type of game offering).

In the **Poker** tab, poker buy and poker sell analysis based on profit is shown. The histogram between poker recency, indicating frequency of use, is also plotted; The **Casino** tab shows the scatter plot for casino mean stakes and casino mean winnings. The **Games** tab shows the scatter plot for game mean stakes and game mean winnings. The **Supertoto** tab shows the scatter plot for supertoto mean stakes and supertoto mean winnings. All of these tabs can be filtered by gender and continents. The **Demographics** tab shows the bar plot for continents and gender. As you can see, Europe has the considerable majority of players, followed by Asia. The number of males are also greater than female.

Conclusion

Overall, the general conclusions per section that can be drawn from the data are that the majority of users are coming from European countries and are male. Section specific-conclusions include:

Sports Gambling:

- Majority of users are between the age of 25-55
- The average stake and average winnings for sports betting was mainly between $0-250
- Most of the users were active within the past 10 days (as of Sept 30, 2005)

Poker:

- Most of the users are making a loss and those who do profit experience only minimal gains
- Most of the users were active within the past 10 days (as of Sept 30, 2005)

Other Games:

- Winnings for casino, suptertoto and other games are again minimal if any and the majority of these users are not repeat/habitual users