Data Analytics

Project - 2019/2020

Designed by: Ester Bernadó

Contents

Introduction

The players of a mobile game have been tracked for a period of time. The purpose was to understand the behavior of these players and the monetization of the company. Data have been tracked for several days. The file **tracking.csv** shows a summary of the actions of the users. Every column specifies:

1. Date: date where the event happens.
2. UserID: identification of the user, which was assigned at installation.
3. Action: action or event that the user is doing: "Install"/"Session"/"Uninstall".
4. Value1: additional information attached to action "Session". NA otherwise.

5. Value2: additional information attached to action "Session". NA otherwise.

Only three actions are tracked:

1. Install: the player installs the game.
2. Session: the player plays the game. Value1 is the time spent in that session (in seconds) and Value2 is the value of the purchases, if any, during that session (0 otherwise). If the player plays several sessions per day, there is only one event registered, which stores the total amount of time spent during the sessions and the total amount of the purchases.
3. Uninstall: the player unistalls the game.

File **users.csv** describes the characteristics of the users:

1. UserID: user identification (a unique number assigned to every user).
2. sex: F(female)/M(male).
3. age of the user.

The designers of the game have been asked to improve the monetization of the game. That's why they need to understand how the players are behaving, before introducing new features into the game. As analysts, you need to provide the answers that the designers have, based on the information provided by the data tracked for more than 3 months.

---

1 General overview of the data tracking

To have a general idea of what has been tracked and for how long, check the datasets provided and answer to the following questions:

1. How long is the tracking period?
2. How many different users have installed the game?
3. How many different users have played the game during the tracking period?
4. How many users have uninstalled the game?
5. What is the average of the number of sessions played by the users during the tracking period?

Perform all the required analyses (with plots and tables) and provide a brief explanation afterwards where you answer these questions.

---

2 Demographics of users

---

1. What is the distribution of users that are playing the game, in terms of sex and age?
2. What is the distribution of age of female players and male players?

Answer these questions by providing nice and interpretative plots. Also comment these plots briefly.

---

3 Metrics of Player Population

The designers of the game need to know the metrics of player population: number of installations per day, churn values (uninstallations per day), and other well-known population metrics such as daily active users and monthly active users. Follow this guideline to answer their questions.

Tip: In all questions, you need to provide the required plots and tables and also a written explanation of what needs to be observed in each case.

3.1 Installs per day

How many installations are there per day?

1. Draw a plot that shows the total number of installations per day. Also draw a line that shows the average of installations. Include a legend describing the two curves.
2. Draw a second plot where you show the same plot as before, together with the number of uninstallations per day.

Tip: a function called "group_by" and "summarize" can be of help. These belong to package "dplyr".

3.2 DAU

How much is DAU?

1. Compute DAU (Daily Active Users) along the days of the tracking period and show a plot (x axis of the play represents the dates and y axis the value of DAU every day).
2. After you plot DAU, draw the same plot, but now with two curves: one with DAU per day and the other one with installations per day. Thus, we can evaluate how DAU is related with the number of installations.

3.3 MAUU

How much is MAUU?

Compute MAUU for each of the months of the tracking period and show it graphically.

3.4 DAU/MAUU ratio

How much is the DAU/MAUU ratio?

1. Compute DAU/MAUU per day of the month.

2. Compute also the average DAU with respect to MAUU which will result in a single ratio per month.

Tip: Consider that MAUU is different for each month. Thus, each DAU should be divided by the MAUU of the respective month.

4 Gameplay Metrics

Now, the analysis looks at the gameplay metrics: number of sessions per user, play time of sessions, etc. Answer the following questions.

4.1 Number of sessions per player

Compute the average number of sessions of the users. Show a histogram and/or boxplot and explain the result.

4.2 Play time

How much is the average time spent by the users in the playing sessions?

The dataset shows the time of each session of every day and every user. One way of looking at the time spent by user is to compute the average gaming time of every user (along the different sessions that each user plays) and the standard deviation. Compute this average and standard deviation. Then, plot the distribution of gaming time of the users as a histogram. Interpret the results.

4.3 Elapsed time between sessions

Compute the average time between sessions of each user and draw a boxplot or histogram. Interpret the results.

4.4 Retention

Compute the retention at day 7, day 14, and day 30. Draw a plot that shows these retentions along the days.

Tip: Since you will need to compute retention several times, the best way is to define a function that does this and that receives the necessary parameters. This is the header of the function:

Retention <- function(DS, dR)

where **DS** is the dataset, and **dR** is the retention that you will compute (D7, D14, D30). The function returns a vector with the specified retention computed along the days of the tracking period.

5 Game Industry Metrics

Now, the designers aim at investigating the common game industry metrics, i.e., the metrics that are related to monetization. Follow the guidelines provided. As always, you need to provide nice plots and tables, as required, and also explanations of the results obtained.

5.1 Life Time

Compute the average life time of the users. Assume that the users that didn't uninstall the game are still active. Thus, compte LTV of those that are not active any more.

5.2 Daily Revenue

Compute and plot the revenue per day and per month. Also compute the total revenue of the tracking period.

5.3 Conversion Rate

Compute the conversion rate of the users of the game. Explain the results.

5.4 ARPU and ARPPU

Compute ARPU and ARPPU. Explain the results.

5.5 LTV per user

Compute LTV of every user and the average LTV of the users of the game. Since active users can be still contributing to the game, consider only those users that churned. Draw the distribution of LTV per user and also compute the average LTV.

6 A/B Test

The players during this tracking period have been used two slightly different versions of the game. Users with ID starting with "ID1" used version 1 of the game, and users with ID starting with "ID2" used version 2. The designers are wondering whether there are diffeerences between these two versions of the game in terms of retention at day 14, and in terms of playing time. Compute an A/B test as follows.

6.1 Retention

Using hypothesis testing, compute whether there are differences between the two versions of the game, with respect to retention at day 14.

Constraints: You are not allowed to use R functions that compute the hypothesis testing directly. You need to compute it by yourself and after that, if you want, you can check if your result is correct comparing with R functions. After you do your analysis, you should interpret it.

Tip: You can use function Retention that you developed above.

To apply the hypothesis testing following these steps:

1. Which type of inference is this (one sample/two sample inference)?
2. Write the null and alternative hypotheses
3. Write which method is the most appropriate and why
4. Compute the hypothesis testing. Justify whether it is one-tailed or two-tailed.
5. Compute the p-value
6. Interpret the results.

6.2 PlayTime of users

Can we conclude that the average play time of users that played version 1 of the game is greater than those that played version 2? To answer this question, you need to compute the average played time of each user in version 1 and version 2. Once you have the two samples, apply the hypothesis testing that is more appropriate.

As the previous section, to apply the hypothesis testing following these steps:

1. Which type of inference is this (one sample/two sample inference)?
2. Write the null and alternative hypotheses
3. Write which method is the most appropriate and why
4. Compute the hypothesis testing. Justify whether it is one-tailed or two-tailed.
5. Compute the p-value
6. Interpret the results.

7 Clustering

Finally, designers are wondering whether there are different types of users. If so, they can adapt features of the games to each user profile.

To analyse whether there are different profiles of gamers, a clustering algorithm will be applied. The features that the designers think as useful to identify profiles of users are: sex, age, average playing time per session, amount spent by the users, average amount spent by the user with respect to the number of sessions, and number of sessions of the user. Using these features for every user, apply a clustering algorithm and investigate which kind of clusters arise. Try with several number of clusters (at least k=2 and k=3) and investigate the results. Then, show plots and interpret the results so that the designers can understand whether there are different types of profiles and if so, which ones. As analysts, also investigate the metrics of quality of the clustering solution. Provide explanations to your results.

8 Discussion and Conclusions

Write the conclusions of your analysis. What can be learnt from the data? A good guide to writing this discussion is to focus on:

1. What knowledge can be extracted from the data? You can structure your explanation on the different dimensions of the analysis: player population, game play, monetization, etc.
2. What issues, if any, need to be investigated more deeply?
3. What recommendations can be extracted that can be useful for the designers of the game?

9 Grading criteria

As analysts, you need to provide a report of the game analytics which is useful to the designers of the game. This means that the report should look as a professional report, not as a simple "academic" work. Some things that you need to consider to have a profesional document are:

- Every plot needs to have its own title, and labels in the x axis and y axis. If there are several curves in the same plot, you need to provide also a legend.
- You may consider using library ggplot2 to print nice plots.
- If you need to show several values, a good idea is to plot a table. Look at the function kable, package knitr, which formats nice tables easily.
- You need to provide explanations (briefly) related to the plots and tables that you are showing. Your report is more than plots, it is about interpreting the figures of these plots.
- The report should be a PDF or html document. To do that, you can use the .Rmd template that is available at ecampus. If you want to have PDF files, you need to install latex packages. Look at the documentation for installing these packages and use them with Rmarkdown: https://bookdown.org/yihui/rmarkdown/installation.html You can also look at information in: https://bookdown.org/yihui/rmarkdown/pdf-document.html
- The document needs to include the R code and the outputs (graphs, tables and explanations) which are required at each section.

Grading of the activity will consider:

- Accuracy of your plots and tables, and degree of adequacy of your answers to the questions posed.
- Quality of the R code.
- Quality of the report (plots, explanations).
- Professional report, as explained before.
- Conclusions and interpretations extracted.