BACHELOR' S THESIS / BIOMEDICAL ENGINEERING 2022

Analysis of the Exploration-Exploitation trade-off in the Koa Foundations app usage

Laura Sánchez Ruiz

upf. Universitat Pompeu Fabra Barcelona

Analysis of the Exploration-Exploitation trade-off in the Koa Foundations app usage

Laura Sánchez Ruiz

Bachelor's thesis UPF 2021/2022

Thesis supervisor(s):

Dr. Rubén Moreno Bote, (UPF Department of Information and Communications Technologies)

Ph. D. Giovanni Maffei, (Senior Applied Scientist, Koa Health)

Acknowledgments

I would like to express my gratitude to Rubén Moreno for making this collaboration with Koa Health possible and for all the knowledge, assistance and feedback received throughout the thesis development. I want to acknowledge Koa Health as well for proposing the project and providing us with the dataset. Special thanks to Giovanni Maffei for the guidance, his valuable aid and advice, and for providing support during all these months, even after leaving the company.

Summary/Abstract

Decision-making is a fundamental human behavior. In our day-to-day lives, we are constantly coming across the problem of choosing among multiple options, which we aim to solve by maximizing the outcomes. To achieve this, the process of making a decision involves the selection of a course of action among all the possibilities in order to stumble upon the best alternative, which becomes especially critical when the decision-maker's resources are limited. In such cases, the usage of optimal strategies for evaluating the options becomes key for the agent to succeed with the task. The decision of the path to take is usually not straightforward, and involves dilemmas such as the exploration-exploitation trade-off, in which the decision-maker agent is faced with deciding whether to sample new options to gain new knowledge (exploration) or instead choose options already sampled and known to perform well (exploitation).

A specific case of study in which such aforementioned tasks take place is in the users' utilization of the Foundations application. This mental-health mobile app offers a wide catalog of activities, grouped into modules, and does not initially provide users with any explicit guidance or recommendations. In this work, data generated by users using the app has been analyzed by tackling the exploration-exploitation strategy framework, ultimately aiming at relating it to users' engagement. This is done under the premise that they aspire to find the activities and modules that work best for them. To that end, in the beginning, they are expected to tend to explore the app more, but then, as their time on the app goes by, a transition towards any form of exploitation conduct might appear. As part of the objectives, the identification of behavioral features that may correlate to the users' engagement with Foundations has been sought.

Keywords

Data analysis, data science, decision-making, Exploration-Exploitation dilemma, mobile app, user retention, user engagement.

Preface or prologue

This thesis has been developed in the Theoretical and Cognitive Neuroscience Lab, a UPF research group led by Rubén Moreno, in collaboration with the startup Koa Health, whose main product is the mental-health mobile application Foundations.

User retention in applications is a broadly spread concern in the mobile app marketing field. In some cases, retaining users may be thoroughly more difficult than attracting new users, which is the reason why strategic business decisions to boost users' commitment are crucial for app success. For example, two recommender systems are already implemented in Foundations, aiming at influencing users to improve their experience. However, a generalized user's poor engagement with Foundations has been recently identified by Koa Health, resulting in a low user retention rate. This has been thought to might be related to users feeling overwhelmed by the excessive amount of content offered, in the form of +200 activities grouped into 9 modules. In addition, the fact that they do not receive any explicit guidance for exploring the app, and thus each of them having to decide their own way, can be a further affecting factor.

A particular way to provide well-founded support in those decisions is through the analysis of how already existing users make use of the app, done employing data disciplines such as data science and data analysis. In this work, techniques from those fields have been applied with the purpose of providing insight into how users are making use of Foundations and how can their adopted behaviors ultimately determine their engagement. More specifically, the analysis framework has been approached from the perspective of a well-known cognitive dilemma that can be identified in the decision-making processes faced by users when deciding how to use the app, the exploration-exploitation dilemma. The final objective of the work consists in relating exploration-exploitation user conducts to engagement, in order to establish a correlation that may indicate potential to predict and improve engagement levels. This way, Koa Health could take action to incite users to behave in the most appropriate way such that their engagement is as much maximized as possible.

Index

List of figures

List of tables

1. Introduction

1.1 Decision-making

Making decisions is a fundamental human behavior. Daily life is made up of a chain of an infinite number of choices. From which degree to study or company to invest in, to the clothes to wear, we are constantly faced with the problem of choosing among options. Behind every action we take, there is an underlying decision, either more conscious or unconscious, behind which there are reasoned causes that are worth to analyze. In the case of humans, these are the product of complex neural processes influenced by many other external factors, and that result in choice strategies that have been identified, analyzed and studied by the decision-making discipline.

Decision-making studies the problem of how a decision-maker agent chooses an option between several alternatives. Three of the fundamental elements of this process are the agent, the available options, and their environmental context. Decision-making tasks involve the selection of a course of action from among two or more possible alternatives to arrive at a solution for a given problem to solve [1]. Such course of action is taken to evaluate the quality of the options and can be influenced by the application of multiple possible search strategies. A key aspect of the process is the way options are evaluated and ranked, which requires a formal definition of what determines an option to be better or worse, strongly dependent on the agent's preferences or criteria. Another critical point is the precise determination of the task, i.e., the precise objective. Another added difficulty is that, eventually, the number of options may exceed those that the agent is able to evaluate due to a limitation of resources, either cognitive or environmental ones, for instance, the agent's motivation or available time.

This research field is hugely broad, and its study involves various disciplines. In the last decades, much of the focus has been placed on mimicking decision-making processes by means of computational and mathematical simulations, using tools such as machine learning techniques. All of these have made it possible to develop formal models, aiming to replicate humans' decision-making behaviors. This way, many decision-making processes have been modeled and listed in the literature, driven by the willingness of answering different research questions [2]. The main reason decision-making processes are studied by multiple disciplines is their inherent complexity, which makes it possible to consider multiple dimensions. The complexity linked to the dimension factor lies in, for example, the fact that when having to make a decision, other prior underlying decisions regarding how to gather information from the alternatives to end up making that choice are needed. Those decision-making processes are needed for evaluating the options to learn about them, and then use that gained knowledge to make further decisions. Such evaluation processes usually involve taking strategies to optimally determine the best way to make the final choice, which in turn commonly implies the appearance of well-known dilemmas and trade-offs, especially when resources to evaluate all options are limited.

A concrete example of a well-defined and studied dilemma in the decision-making theory is the exploration-exploitation (EE) one. Such dilemma arises when the resources to evaluate options are limited and is likely to appear in many decision-making processes of real cases of study together with other ones such as the breadth-depth (BD) dilemma [3]. The EE dilemma implies a trade-off in the behavior of the decision-maker when having to face the task of evaluating the candidate options, which makes the existence of an optimal balance between those two possible searching strategies (exploration and exploitation) likely.

1.1.1 Exploration-exploitation dilemma

The exploration-exploitation dilemma is a well-known and broadly examined problem in the community of decision-making science. This dilemma appears when performance is to be maximized, exposing a trade-off between exploring new options, with the purpose of gaining knowledge, or otherwise exploiting those others that are already known to perform well [4][5].

A usual case exemplifying this trade-off is the machine reinforcement learning problem known as the multi-armed bandit problem, which comes up in situations in which, when aiming at maximizing the expected gain, there is the need to allocate limited resources among the competing options to evaluate them [6]. The most classic example representing the bandit problem is the casino slot machine setting, in which one must obtain the maximum earnings by gambling on the machines, having each of them a different and totally prior unknown behavior [7]. In this case, the limiting resource to be allocated would be money, as well as the reward to maximize, and the path to determine the best machine is composed of sequential rounds that consist of betting in a machine and receiving immediate feedback, from which one gathers knowledge. The dilemma appears in the sense that, once identified a machine that performs well (i.e., the one that is giving the most earnings), one is faced with wondering whether another machine may perform better. This way, the decision-maker is confronted with having to adopt a course of action taking into consideration the two possible bents: either to continue sampling a currently and seemingly optimal option or, instead, to assess other potentially better alternatives. In principle, presumably, it seems deducible that one must not allocate all resources to a good machine (exploiting it) but instead should maintain a certain degree of exploration just in case a better option is still needed to be identified.

1.2 Case of study: Foundations mobile app

A particular example in which there is an underlying decision-making task involving dilemmas such as the EE and BD happens in the Foundations application, a product of the startup Koa Health. Foundations is a mental health app that provides a catalog of interactive activities and programs that aim to improve users' mental health and wellbeing. These activities and programs, based on clinical evidence, are offered as tools for helping users learn about how to work on their mental wellbeing. All these tools are disposed grouped by a total of 9 categories (or modules) in the interface of the 'Library' app section and structured in a way such that users are free to explore the full catalog without any explicit guidance.

Foundations is designed to tackle a wide range of user profiles, from people with sleeping disorders to those suffering from anxiety. In that sense, it is said to be a generalist app, as it offers a considerable amount of content to all users. On the one hand, the fact of having more than 200 activities is a positive aspect, making the app rich in content and therefore potentially useful for a higher portion of the population. In addition, giving users the possibility to explore the offered options and make their own choices while being proactive in trying to improve their health has been proven to be motivational and beneficial for them [8]. On the other hand, however, providing users with a too wide range of activities together with a lack of explicit guidance in the exploration process might hinder the task of finding the activities and categories that better fit them. In multiple studies, it has been observed that not having a small enough number of options to be able to evaluate all of them, leads subjects to act heuristically to reduce the alternatives [9][10][11]. This situation has been identified as a possible cause of the decision-maker's poor engagement with the task, showing higher compromise when being faced with a limited number of alternatives that are plausible to be tackled by the agent. This is strongly correlated with the decision-maker capabilities, which are usually limited by multiple factors such as, in this case, users' available time or motivation to explore the app.

Recently, Koa Health identified a poor generalized users' engagement with Foundations, as up to 70% of them uninstall it after a month of usage. Low retention rates constitute a generalized issue in the ubiquitous mobile apps marketing, a fact that evinces the need for understanding in which manners and contexts do users engage or not with such apps in order to achieve the product's success. A strategy to provide insight into apps' usage and extract meaningful information that allows to give support in making subsequent business decisions is to analyze the users' behavior through data analysis and data science methods (such as machine learning techniques), which can even further allow for predicting users' retention and level of engagement. Several studies tackling such problem have been reported in the literature [12][13][14][15].

In this case of study, because of all the previously exposed reasons, it has been thought that the excessive extension of the offered catalog might be the cause of the general low engagement level. In this context, the relevance of an analysis of the way users make use of the app in terms of how they take the sequential decisions to test the offered content becomes clear. Specifically, an interesting way in which this decision-making problem can be approached is from the perspective of either the BD or the EE dilemma. Both appear when resources are limited, but the first one is rather identified in the early exploration phases, as it consists of a trade-off between allocating resources to many or instead only to a few of the alternatives to evaluate them when there is neither prior knowledge about them nor immediate feedback. In this work, the analysis of the users' behavior is approached from the EE dilemma framework. In parallel, a supplementary study addressing the same project from the BD perspective has been conducted by Pujol [16].

- **Decision-making and EE dilemma in Foundations**

For setting out the dilemma in this scene, it is crucial to appropriately define the identified decision-making task and its key elements. The lack of explicit guidance in Foundations evinces the need for establishing a strategy to proceed to test all the available content, such that the main implicit decision-making process corresponds to the selection of a course of action to test the app content. Users do so by making sequential decisions that shape their search strategies, which have been assumed to have the furthest goal of finding the tools that suit them best, i.e., those that they enjoy the most. This way, the final decision-making task underlying this decision process consists in deciding on which content is the preferred one, that which best works in each case.

In this case, except for the decision-maker, which corresponds to each of the app users, a precise definition of the other essential elements of the decision-making process is not as straightforward. That is why some of them have been deducted, resulting in several made assumptions, while the identification of other ones is even part of the objectives of this work. Those elements include the following:

i. **Options**: the alternatives to evaluate and choose between are the activities offered in the app, which are grouped forming categories.
ii. **Agents' resources**: identified as their time and/or motivation to play with the app. Taking into account that the number of activities is over 200, resources have been assumed to be limited, even if not being quantifiable, what intrinsically implies that users are not able to test all alternatives, thus must reject some and choose which ones to evaluate and how.
iii. **Task**: assumed to consist in, by evaluating the options, getting to determine the content that better works for each deciding agent. There are two key details to clarify about this. The first is the criteria that rule the quality of the options, which is clearly subjective (depends on the user preferences). Furthermore, the fact that these subjective rules can (and, actually, are likely to) be dynamic over time, confers an added complexity to this definition. The second aspect to consider is about the users' objectives. Based on the prior assumption of them aiming at finding the best content, and considering the characteristics of the app as for its layout design, a question arises

as to whether users might be looking for either the best activity or the best module instead. This last question, linked to the whole specific definition of the task, has been addressed in the work as one of the unknowns in the hypothesis.

In this framework, the EE dilemma comes into play in the following manner: once a user has found the activity or module that seems to work for him, should he restrict himself to that content, exploiting those options which are functioning? Or should he continue exploring the application just in case there are other options that may work even better at the cost of the possibility of wasting resources? Taking the two levels of content organization (activities and modules) into consideration, the following different possible exploration-exploitation scenarios arise:

- **Pure exploration**: exploring activities of different modules.
- **Balancing exploration-exploitation**: exploring activities within the same module, and so exploiting modules.
- **Pure exploitation**: exploiting activities.

This thesis seeks to identify these kinds of conducts in the behavior of users when managing the app. Moreover, those EE practices might vary over time and evolve towards some specific trend as the users make more use of the application. In that case, users at early stages of their stay would be expected to mostly explore, and tend towards any conduct involving exploitation as their time in the app went by.

2. Hypothesis and objectives

2.1 Hypothesis

The primary hypothesis of this work is that exploration-exploitation behaviors can be identified in the users' employment of the Foundations application. The trade-off in this case of study is hypothesized to appear in the process of the search for coming up with either the best activities or the best modules, understanding them as the ones that best fit each individual.

It is hypothesized that users experience a first greater exploration phase that evolves into a major tendency towards the exploitation of activities and/or modules already known to work for them, which might even end up leading to a balance between such exploration and exploitation conducts.

Finally, an ultimate premise is that there may be a correlation between those EE behaviors in the app usage strategies and their app engagement.

2.2 Objectives

The first objective is to determine the users' behavior and search strategies when testing and evaluating the multiple presented options, from the exploration-exploitation dilemma framework, to validate or reject the hypothesis.

The second major and furthest goal is to be able to correlate the aforementioned behavior strategies to the users' commitment to the app, recognizing optimal strategies. For this, engagement metrics will have to be decided and it will be necessary to identify dissimilarities between users with different levels of engagement. The main aim of finding such correlation would be to allow for predicting users' engagement depending on their shown EE conducts, especially relevant at early stages of their stay. Consequently, business strategies could be

established to promote a more efficient app utilization that lastly leads to an improved user retention rate and app engagement.

All of this is meant to be done through the analysis of the data collected by Koa Health, employing data science principles and data analysis techniques. The overall objective is to efficiently use the existing methods and techniques to carry out a full data science process. To do so, some specific objectives encompass the following:

1) Determine the most relevant raw data for the analysis specific approach
2) Extract and visualize the determined information of interest contained in the provided dataset
3) Perform an overall analysis of the app usage
4) Establish lines of action for the specific in-depth analyses, one for each of the two main objectives and their underlying hypothesis: EE identification and engagement correlation
5) Interpret all the results and extract conclusions useful from a business perspective

3. Materials and methods

The main tool employed for the thesis has been Python 3.9[1], by means of which the determined data of interest for testing the hypotheses and achieving the objectives has been extracted, visualized, and interpreted. To do this, the principal methodology followed in this work has been the application of data science principles, combined with data analysis techniques, all approached from a Business Intelligence (BI) scene. Briefly, the term BI refers to all those tools and procedures used by companies to transform information extracted from acquired data into knowledge that is useful from the business perspective. This allows to, in a strategic way, provide well-founded support when it comes to making decisions that aim at improving a product or service. To do so, the process involves lots of techniques belonging to multiple domains of the study of data, such as data mining, science, and analysis.

3.1 Data science

Data science is a scientific interdisciplinary field that consists in analyzing vast amounts of past data to extract and interpret information that allows identifying trends or patterns, aiming to make future predictions. This discipline focuses on the establishment of the proper means to accomplish the extraction of those business valuable insights, done by making use of multiple tools and methods such as machine learning or alternative analytical methods. In a corporate and BI context, its ultimate goal resides in the ability to identify business opportunities on the basis of the extracted information. Although the way the data is approached may be subject to slight variations, every data science project consists of a lifecycle that generally involves the following sequentially interconnected steps [17][18][19]:

i. Framing the problem: every process starts from understanding the project's specific requirements from a business perspective. This is followed by posing an interesting question, defining the objectives and problem, and converting it into a data mining problem through the determination of the strategy to follow to tackle and fix it.

ii. Data collection: linked to one of the core disciplines in data science, the second step is focused on the non-trivial determination of the raw data which is potentially relevant to be analyzed in that particular context. After that, it is proceeded to gather that determined accurate data in a specified format, which will be usually raw. Such discipline is data mining, whose function is to extract valid, new, and useful information from raw data to find hidden patterns in them.

[1] https://www.python.org/

iii. **Data exploration and preprocessing**: after obtaining the raw data, it needs to be understood to identify whether it is appropriate for answering and solving the posed question and problem. For that purpose, a raw data overall view is required, done through a first exploratory data analysis (EDA). After understanding the data and performing the EDA, the raw data needs to be prepared and cleaned for it to be treatable and easier to handle.

iv. **In-depth data analysis and modeling**: data preparation and cleaning are followed by data manipulation. This is mainly about feature engineering, i.e., coming up with new meaningful variables that can be extracted from the available raw data and are interesting to inspect and analyze in-depth. Depending on the problem being tackled, one determines the pertinent analyses to perform in between the three main types of data analytics: descriptive of what happened in the past, predictive of future events, or prescriptive, which is a combination of both strongly correlated to BI.
Finally, the determined pertinent analysis is performed, with the objective of generating a model, either descriptive or predictive. The former kind of modeling is based on recommendation systems and clustering algorithms to extract information about which services are interesting, whereas the latter aims at predicting future expectations through classification and regression algorithms.

v. **Analysis results visualization, interpretation and conclusions**: the final step after obtaining the results by deploying the model is to interpret them. Such interpretation must be followed by the company users' feedback gathering and ultimately to the results linking to business opportunities.

3.2 Raw data

The materials of this study mainly consist of the raw data supplied by Koa Health, corresponding to information regarding how users navigate through the app. It is important to remark that two different datasets were received, such that naturally the determination of the analysis approach and strategies to take with each version of the data was different. However, some of the initial ideas could be applied to the second and final analysis.

The full dataset comprised three JSON files with varied information about the users. The main file includes data about their experience with the app collecting a record of the events (logs) generated by users when making use of the app. The other two supplementary files consisted in a mapping between activities and modules, and a collection of users' details about their whole stay in the app, such as their estimated spent time or number of sessions.

The logs file gathers every action taken by users, all saved as events with some specifications each. Five types of events are defined: impressions, screen-views, taps, starts of activities, and finishes of activities. The first ones were neglected from the data due to their unreliability, detected by the company itself. Screen-views events are created when it is accessed to a particular section in the app, and taps events when pressing on any clickable app site. Events have some shared characteristics (e.g., the time at which they take place or the user by which they are generated) and other hallmarks that depend on the specific event. Each log is saved accompanied by its own specifications, both for shared properties and other event-dependent characteristics such as the activity ID, indicated only in cases of events involving activities.

event	view	element	created_at	activity_id
tap	library	see_all	2021-06-13 16:28:38.893000	None
screen_view	unit_reading	None	2021-10-19 19:44:09.904000	71621f00-a161-4b8…
screen_view	focus_area_detail	None	2021-10-19 19:50:32.510000	None
start_activity	activity	None	2021-10-19 20:06:26.012000	cdbb4000-deb0-4b4…
tap	activity_outro	good	2021-06-21 20:46:00.172000	a1c2456c-9bbb-4f0…
screen_view	activity_outro	None	2021-06-21 02:14:35.711000	b123aec9-f580-45e…
tap	activity_outro	good	2021-05-01 01:14:08.199000	17a99f89-8c11-438…
screen_view	programme_detail	None	2021-06-02 12:23:35.293000	None
screen_view	activity	active	2021-06-03 22:27:24.900000	None
tap	quiz_activity	answer	2021-06-03 20:11:37.091000	5fa57d00-e0fc-4af…

***Figure 1**: Illustration listing some examples of events, extracted from Python.*

As stated above, we were given two raw data versions. Initially, the dataset information corresponded to the app usage during a single month, November 2021. Later, due to analysis requirements, a larger period of time was requested and conceded. In the end, the extension of the employed data comprises April 2021 to April 2022.

3.3 Data reading and handling

The dataset has been read, manipulated, and analyzed with Python as the programming language by means of Spyder, an integrated development environment specifically designed to be significantly useful for the data science and analysis disciplines, due to the incorporation of some tools and libraries such as pandas [20]. Many functions of those libraries have made it possible to read, process, and visualize the raw data as well as the extracted features of interest.

The data in the files when read with Python is in the form of DataFrame objects, which are 2D labeled structures with a tabular look collecting positionally accessible data. In the logs file, the DataFrame object rows correspond to every single event, each detailed by its properties in the corresponding columns, specified or not depending on the event type (see Figure X). The definitive file (without considering impressions) gathers a total of 1,850,724 events generated during a whole year by 7,777 different users.

3.4 Exploratory data analysis

In consistency with the data science steps, after having understood the project from the business side, having identified and framed the problem question, and before starting with either the data manipulation or the analysis itself, the way to proceed is to determine the data to be collected, i.e. that which is potentially relevant for the analysis. In our case, initially, this step was directly performed by Koa Health as the raw data was already provided. However, this prime dataset was mainly used for performing a first general analysis while the definitive approach details were being defined. After that, as mentioned before, the determined final approach raised the need for obtaining a wider period of time of the information.

When finally obtaining the definitive raw data, it was needed to determine which events were potentially useful for the thesis objectives. In this case, the selected logs were the start activity type, as those were the ones that were assumed to correspond to users testing the offered options. In addition, other further variables identified to be relevant for a first overall analysis were already obtained from the raw data. For each user, the following variables were extracted:

- **Onboarding date**: corresponding to the date of the onboarding event, that is of course the first generated event.
- **Total generated events**: obtained by counting the total number of logs generated by each user.

- **Total days in the app ("stay days")**: its definition was tricky as there is not an event that indicates the users' offboarding (i.e., when are they leaving the app). Initially, to estimate this value, the dates of the onboarding and of the last generated event were used. However, this latter led to problems and, after being redefined, the "offboarding" date was finally obtained by subtracting the onboarding dates from the dates of when the last activity was started.
- **Days visiting the app ("visiting days")**: from the total staying days, how many of them does each user generate at least one event.

The provided data together with the just exposed extracted features were used to perform a first exploratory data analysis mainly through graphic representations. The objectives of the EDA were to get a general view of how users make use of the application by describing the data, as well as to end up establishing and tuning the hypothesis and concrete objectives. This way, it was also useful for determining the appropriate subsequent analysis strategy and basis, including among others the determination of some filtering specifications. It is important to mention that part of the EDA was repeated after performing the data filtering.

3.5 In-depth data analysis

3.5.1 Basis of the analyses

After the EDA, linked to the established hypotheses and objectives, two main lines of action were differentiated: one first oriented to the identification of exploration-exploitation conducts, and another focused on associating such behaviors with users' engagement. Different analysis approaches were taken in each, from the data filtering to the interpretation of the results. Nevertheless, a common trait between the two workflows is the types of generated events that were addressed, which as already mentioned were mainly the starts of activities.

- **Exploration-exploitation analysis**

The idea of this approach has been to analyze the evolution of the users' behavior when using the app as time goes by, to identify possible differences that can be interpreted as the adoption of exploration and exploitation-like dynamics. As exposed in section 1.2, this transition might appear at the level of activities or modules. To do this, the basis of the EE analysis is to compare data of the same users in different and consecutive time periods. The establishment of those time periods has been mostly arbitrary, as in the case users showed exploration-exploitation conducts there is no prior way of knowing when they will appear. In fact, that is an intrinsic objective of the thesis, and the most probable is that it is a user-dependent thing. In trying to solve this, 6 groups of users depending on how many total days they are in the app have been established (see Table 1 on Results) based on the premise that depending on this variable, they will be prone to expose EE behaviors at a distinct rate. This way, some assumptions have been required with regard to both the user groups and the time periods to be inspected. Based on such premise and after several trials of a few possible time periods, finally, the considered ones have been 1 week and 2 weeks.

- **Engagement analysis**

This second analysis has been focused on trying to relate users' app usage, as regards exploring or exploiting it, to their engagement with the app. The main key needed for doing this is the determination of how to measure such engagement. The selected features for studying long-engagement have been the total number of days in the app and the ratio of days visiting it, which have been combined for creating a scoring system. Both total and visiting days have been normalized by the total days that each user could have been in the app, to take into account that, depending on their onboarding date, each user has had a different number of total possible stay

days. This has been done by dividing the variables by the difference in days between onboarding and the latest date of the dataset.

Of those two features rating engagement, the one recognized as most valuable is the percentage of days visiting the app, which has been assigned a higher weight. The reasons behind this are that, as said, a high percentage of users (~70%) abandon the app after a month, falling the total days in the app too short for measuring compromise. In addition, the fact that a user enters the app a high percentage of the stay days is a greater indicator of engagement, independently of the total stay days (if high enough). The devised equation to obtain the normalized engagement score is represented as:

$$ES = 1.5 \times \frac{100 \times v_{norm}}{s_{norm}} + 0.8 \times s_{norm}$$

***Equation 1**: established formula for the engagement score, used as long-term engagement metrics. The variables v and s respectively stand for visiting and stay days, and subindices "norm" indicate normalization by total possible days.*

After having determined the engagement metrics, 6 other user groups depending on their engagement score were created (see Table 1 on Results). With this, the basis of this analysis has been to compare the data of that different engagement level type users at the same time periods, in order to identify differences in their behaviors that may be affecting long-term engagement.

3.5.2 Data filtering

With the analysis strategy defined, the next step is data cleaning and preparation, i.e., to perform the raw data filtering. The purpose of this step is to make data easier to manipulate by keeping only that information that is useful for the case of study. First, the events which are not of the selected type (start activity) have been filtered out. Then, a selection of the users considered to be useful has been made by dismissing the ones that do not fulfill the following requirements:

- Onboarded between the period of time covered by the data, otherwise it is not possible to know how many days a user has stayed and visited the app
- Started at least one activity in the period of time covered by the data, otherwise they are considered to not hav made enough use of the app
- Stayed in the app for at least 8 days, otherwise it is assumed that one has not had enough time to show the inspected behaviors
- Have visited the app a sufficiently high percentage of days, mainly to discard for instance users that stay more than 200 total days but only enter the app between 1 and 4 days

With this, the final number of selected users is reduced to 919. Regarding the kept logs ("start activity" from the selected users), some of them were detected to be duplicated, an artifact fixed by eliminating the events that were generated by a same user at exact same dates. Furthermore, when merging the logs file to that containing the correspondences between activities and modules, there were a few activities whose mappings were not being specified and thus were neglected for the analysis. Finally, while the analysis was being carried out, whichever data was identified to be missing, was appropriately filled in. For example, when obtaining the total number of activities started by each user, this information for those who did not start any had to be manually added.

3.5.3 Data extraction and visualization

3.5.3.1 Metrics and plots

After having filtered the data and established the user groups, another brief exploratory data analysis was performed. Then, it was time to extract the information of interest for the analysis. Apart from the already obtained features for the EDA (those exposed in section 3.3), a further feature engineering process was done. However, this time the obtained metrics were those that were identified to be pertinent for identifying exploration-exploitation behaviors, so as to directly tackle the first and principal hypothesis. All these variables were extracted from the start events and are as follows:

a) Total number of started activities
b) Unique (i.e., different) started activities
c) Unique started modules
d) Total number of activities repeats
e) Total number of modules repeats
f) Unique repeated activities
g) Unique repeated modules

These metrics were computed per each user and also in turn per the established time periods (both per week and per two weeks). To do this, lists of the users present in each time point were defined, a fact that was dependent on their "offboarding" (last start) date. Equally, other lists corresponding to the user types, both per days in the app and per engagement level (so a total of 12 lists), were also denoted to allow for the separation of data based on this criterion.

- **Entropy**

To end with the EE metrics, an ultimate measurement was used: Shannon's entropy [21]. This concept of entropy belongs to the information theory field of study and is correlated with disorder and dispersion. Entropy has long been employed as a measure of spread within a distribution [22], so that it allows for measuring the variability of the elements within a given distribution, or in other words, it characterizes how much variation can be found within a collection of samples [23]. In this case, it was used to investigate the variability of the started activities and modules. The higher the entropy, the higher the variability of the samples, meaning that the sample allocation is more even, i.e., the user tests more different activities or modules with respect to the total started (higher exploration). Entropy is calculated with the formula in Equation 2, either in natural units if using ln as the log function or in binary units if using $\log_2$ instead [24].

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

***Equation 2**: Shannon's entropy for a discrete random variable X, whose probability is p(x) and stands for the probability of observing a particular sample.*

At first, the entropy metrics was thought to allow for looking at the variability of started activities and modules in a way that it was independent of the total number of started activities. However, entropy intrinsically brings about an important bias to consider, indeed related to the number of samples. The simplest example to understand it is by thinking about the extreme case: for a single sample, the entropy is minimum (namely, 0). Hence, the larger the number of samples, the higher the entropy will most likely be, which intrinsically implies that the higher the number of total started activities, the more probable it is that the entropy is higher, and the other way around. For solving this problem, cases of only starting a single activity have been rejected and subsampling has been applied to the rest. The entropy for each user has been computed by iteratively (x100) calculating the entropy associated to two selected samples, in a way that the samples are randomly taken *but* taking into account their occurrences. In this setting, those samples (i.e., activities) can

be a same repeated activity (minimum entropy) or a different one (maximum entropy), such that the entropy of each iteration becomes binary, reason why base 2 logarithm was used for its calculation. Ultimately, the mean of the 100 obtained entropies per each user is assigned as the resulting corrected entropy.

As regards the types of visual representations generated, they consisted of histograms, density distributions, box plots, and standard line, bar and scatter plots. As a note for the density distribution plots, they were obtained by using the kernel density estimate (KDE) technique, which estimates the probability density function of a random variable [25]. It is analogous to traditional histograms, but unlike those, it enables to better analyze the studied probability distribution, as it can produce a more interpretable plot, especially when drawing multiple distributions [26]. On the other hand, box plots allow for a better visual comparison between levels of a categorical variable by plotting the distribution of quantitative data. They contain three main elements: a box showing the quartiles of the dataset, whiskers which extend to show the rest of the distribution, and points determined to be outliers using a method that is a function of the inter-quartile range [26].

3.5.3.2 Techniques

The techniques used for obtaining those metrics and plots are data analysis and data science programming techniques in Python. Apart from widely making use of libraries such as NumPy [27] or other programming resources for coding variables handling, one recurrent strategy followed to collect all the engineered features has been to conveniently group the data in new created DataFrames. The most meaningful examples of those are:

1. A DF indicating, for each user, general extracted information such as the total number of days in the application, the engagement score, and the groups they belong to (see Figure 2 below).

hashed_user_id	onboarding_date	last_start_date	days_visit	norm_num_days_visit_apu	days_in_app	norm_days_in_app	eng_points	norm_eng_points	days_type	eng_type
aa8e710317d9d…	2021-07-31 …	2021-09-30 …	14	5.10949	62	22.6277	34.9806	51.9732	Type III	Level 2
4165011b506e7…	2021-07-31 …	2022-04-29 …	27	9.85401	272	99.2701	64.3265	94.3058	Type VI	Level 5
7ba834a736090…	2021-08-01 …	2022-03-29 …	26	9.52381	241	88.2784	58.9884	86.8053	Type V	Level 4
2817699a77f20…	2021-08-01 …	2022-04-27 …	121	44.3223	270	98.9011	98.8148	146.343	Type VI	Level 6
6c7c391201fda…	2021-08-01 …	2021-09-27 …	33	12.0879	57	20.8791	69.2947	103.545	Type II	Level 5
c10bcad990d2a…	2021-08-02 …	2021-10-27 …	7	2.57353	86	31.6176	25.3395	37.5034	Type III	Level 1
d9e7764cb41af…	2021-08-02 …	2021-09-21 …	5	1.83824	51	18.75	20.0039	29.7059	Type II	Level 1
be2f974dfb6ac…	2021-08-02 …	2022-04-20 …	117	43.0147	261	95.9559	97.0276	144.006	Type VI	Level 6

***Figure 2**: Example of generated DF object for code programming purposes, containing some extracted data for each user. For instance, at the end, one can see for each case two columns indicating the type of user per number of days in the app and per engagement.*

2. Another DF listing the users specifying in each case all the metrics listed in the previous section (see 3.5.3.1) in each of the determined time periods.
3. Another DF like the previous one but with the entropy measurements in each time period, both those initially obtained and those corrected by subsampling.

Those main working DataFrames were strategically separated or merged if necessary or most opportune for programming requirements. To do all this, as well as to extract the features from the logs DF, functions that belong to the pandas [20] library have been utilized. Frequently used functions comprise *merge* and *concat* to combine DF objects, and *groupby*, which allows for grouping a DataFrame by a specified column. As an example, the simplest case of use was to get the total number of started activities, obtained by grouping the start activity logs by users.

The rest of the metrics (start, finish and repeat activities or modules) were obtained similarly. For the activities and modules entropy calculation, it was first needed to define, for every single user and each of the subsamples, the probabilities of starting each of the different started activities or modules. Intuitively, those probabilities can take values 0.5 and 0.5 if samples are a different activity, or 1 if it is a repetition. Given those probability values, entropy was calculated by applying the statistical function *entropy*, from the SciPy scientific computation library [28].

For generating the plots, functions belonging to the Matplotlib [29] and seaborn [26] libraries were employed. The first is generally used for creating visuals, and the used functions were *plot* and *scatter*. In several of the scatter plots, their data was fitted to a power law by means of the *curve_fit* function from SciPy [28], whose accuracies were determined by means of the coefficient of determination (R^2). Also, a single plot was fitted to a polynomial of degree 1, with the NumPy [27] function *poly1d*. On the other hand, seaborn, which is more specific for statistical data visualization, allowed for plotting histograms (*histplot*), density distributions (*distplot*) and box plots (*boxplot*).

- **Statistical significances**

Statistical significances indicate the reliability of specific statistics, required for the verification of the possibility for the analysis results to be explained by something else other than pure chance. It is usually determined by calculating the p-value, typically considered to indicate statistical significance when $p \leq \alpha = 0.05$ and obtained by means of statistical tests. In this work, the appropriate was to use non-parametric tests, which do not need any distributional assumption [30]. Three different tests were used:

i) **EDA and EE analysis**: as the data to be compared belongs to the same population (i.e., it is paired) under different conditions, Wilcoxon Signed-rank test was used for pairwise comparisons. This test is the analogous to the parametric paired t-test, and tests the null hypothesis that the medians of two matched samples are equal.
ii) **Engagement analysis**: the data to compare belongs to different users (i.e., it is not paired) under the same conditions. The first test used is Kruskal Wallis, which tests the null hypothesis that the medians of independent groups are different, comparing the ranks of the data (rather than actual data values), allowing for the comparison of multiple groups at once. However, as it does not specifically indicate *which* groups are different, if its result indicates statistical significance, a posterior post-hoc test is required [31][32]. The one used was Dunn's test, which compares all groups in pairs.

For the first two tests, p-values were obtained using the SciPy [28] statistical functions *wilcoxon* and *kruskal*. Finally, Dunn's tests were carried out by means of the function *posthoc_dunn* from the scikit library [33].

- **Standard Error of the Mean (SEM)**

The SEM is a metrics of precision for an estimated population mean [34], which in this case was used for measuring how far the calculated sample entropy means of the data were likely to be from the true population means. For doing so, the following equations were used:

$$SEM = \frac{\sigma}{\sqrt{N}}$$

***Equation 3**: Formula for the SEM, where σ is the standard deviation and N is the number of total samples from which the mean is calculated (population).*

The standard deviation is calculated by means of:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\langle H\rangle_i - \bar{H})^2$$

***Equation 4**: Formula for the variance. Here, $\langle H\rangle_i$ corresponds to each calculated entropy per user (result of performing the mean of the calculated entropies per subsample), $\bar{H}$ to the calculated mean entropy for the population (in this case, the mean of the means), and N to the population size.*

4. Results

As a data analysis work, a large number of plots have been generated throughout the project development, such as for determining the filtering or other punctual analysis requirements. Here those final graphs containing relevant findings for the work hypotheses are exposed, being some of the rest gathered in Supporting information, where a collection of the obtained p-values for statistical significance testing in each pertinent case are listed as well.

4.1 Exploratory data analysis

The generated plots for determining the user types can be checked in Supporting information (see Figure SI 2 and Figure SI 1). The final defined groups and some of their details are summarized in the following table:

Group name	**Type I**	**Type II**	**Type III**	**Type IV**	**Type V**	**Type VI**
Days in the app range	8 - 28	28 - 58	58 - 100	100 - 160	160 - 260	over 260
# users	225	152	125	130	167	120
Group name	**Eng. Level 1**	**Eng. Level 2**	**Eng. Level 3**	**Eng. Level 4**	**Eng. Level 5**	**Eng. Level 6**
Score range	25 - 40	40 - 55	55 - 70	70 - 90	90 - 105	105 - 226
# users	145	206	169	211	100	88

***Table 1**: Established groups of users depending on the stay days in the app (upper part of the table) and depending on their engagement level (lower part of the table). For each of the 12 groups, it is shown: the specific given name, range of the corresponding variable used for the grouping and total number of users.*

An additional extracted datum from the EDA that affected the rest of the analysis was that none of the type VI users starts any activity until day 48 (see Figure SI 3a), reason why they were not considered for the EE analysis.

In the following figure, one can see an example of the plots for the total number of started activities metrics.

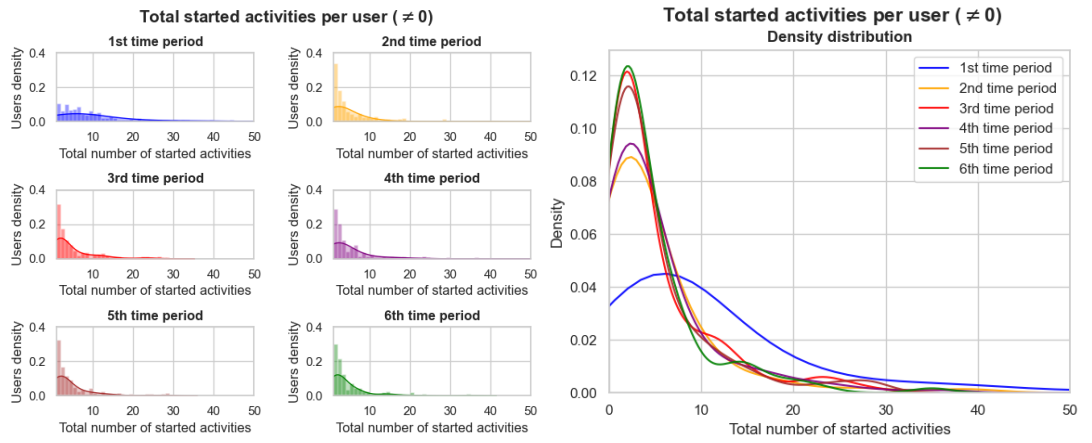


***Figure 3**: Aggregated data of the total number of started activities per user for each time period of 2 weeks, dismissing zeros for better visualization. The considered user types are II, III, IV and V. **(a)** Histograms and density distributions plotted separately per time periods. **(b)** Density distributions per time period in a unique plot.*

In Figure 3 it can be observed how, as time goes by, the density distributions get narrower and concentrate and increase their peak at low values. This is the general behavior observed for all the other metrics as well (see Figure SI 4). Moreover, another shared observation is that the differences between the values of the first time period (namely either 1 week or 2 weeks) with respect to the rest of periods are the most statistically significant, showing p-values between orders of 10^{-15} and 10^{-8}. This is especially emphasized when comparing the first to the second time periods, with p-values around in between 10^{-22} and 10^{-28}.

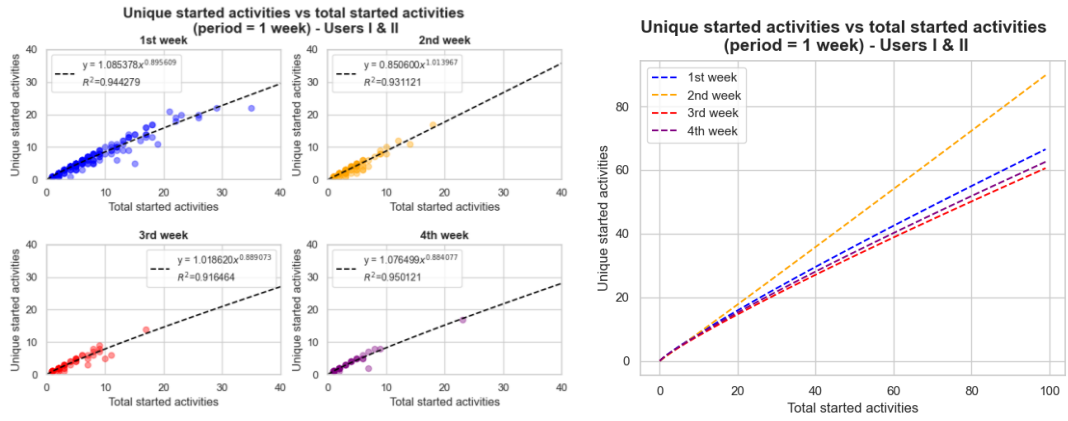
4.2 Exploration-exploitation analysis

The results of this analysis mainly consist in two different types: the comparison of the power law functions fitted to different scatter plots and the evolution of the entropies over time periods per user types (per days in the app). All the exposed plots for the first result types correspond to aggregated data, meaning considering data of multiple various specified user types. A single example of the graphs displaying the data points (scatter plots) and corresponding fits is provided below (Figure 4a).

4.2.1 Activities

In this section one can see the results for looking at the exploration-exploitation behaviors regarding activities. Two plots are shown below, one with the fitting example (Figure 4a) and another one displaying the tendencies of how many different activities users I and II are expected to test from week to week as a function of the total started ones (Figure 4b).

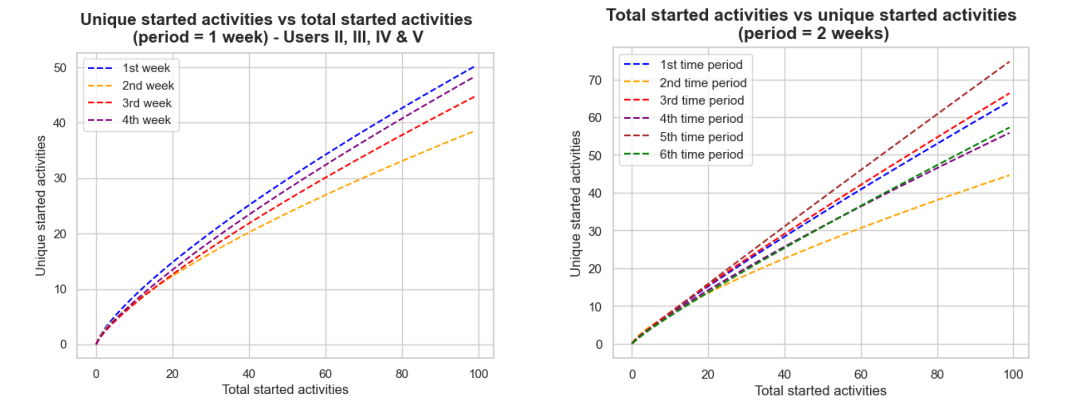
In Figure 4a, one can observe that the generated fits are of remarkable quality, i.e. with high associated coefficients of determination (R^2). All differences are statistically significant, with the exception of those between the 2nd and 4th weeks. It can be seen how the fits for the first, third and fourth weeks have similar exponents around 0.9, whereas the power law exponent for the second week is approximately 1. This indicates a linear relationship, meaning that almost all the started activities tend to be different activities.



(a) (b)

***Figure 4**: **(a)** Scatter plots showing, in an aggregated manner for user types I and II and per weeks, the number of different started activities (vertical axis) as a function of the number of total started activities (horizontal axis), each together with their associated fitted power law function (black dashed lines). **(b)** Power law fits for each time period (week). Associated p-values in **Table SI 1**.*

The following figure shows the same results but for users staying more than 28 days (namely, types II, III, IV and V). Additionally, for those user types, the two time periods were tested:



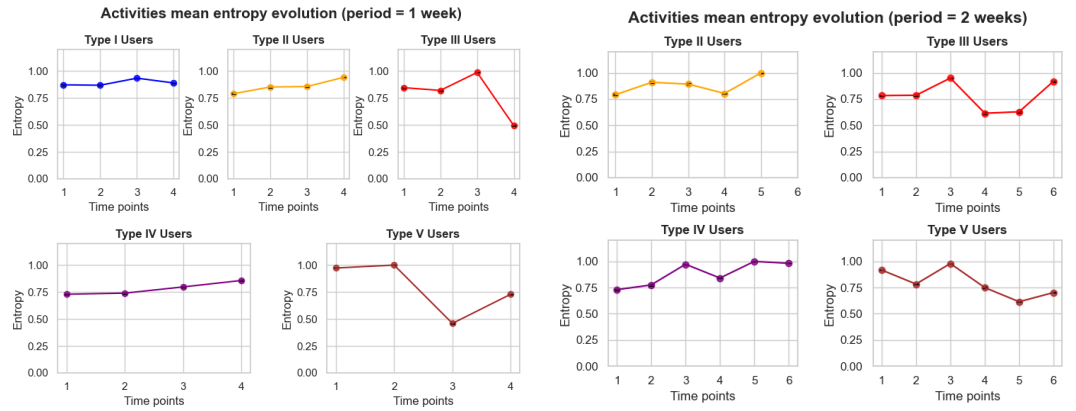
(a) (b)

***Figure 5**: Power law fits for the unique started activities as a function of the total started activities, in an aggregated manner considering user types II, III, IV and V, and per time periods. **(a)** Time period of 1 week. **(b)** Time period of 2 weeks. Associated p-values for both in **Table SI 2** and **Table SI 3**.*

Again, the R^2 values obtained for the fits indicate good fitting quality (see Figure SI 5a and 5b). On the left graph in the figure above, only the 1st week fit is statistically significant against the others, which can be appreciated to be the least flattened. This represents a tendency in the first week for trying a higher percentage of different activities. On the right, the 1st time period is only statistically significantly different to the 2nd and 3rd. When comparing the first half month to the second, one can see that the portion of different activities tested decreases, similarly as from the

first to the second week. After that, in the next two weeks (3rd time period), users seem to return to the initial behavior.

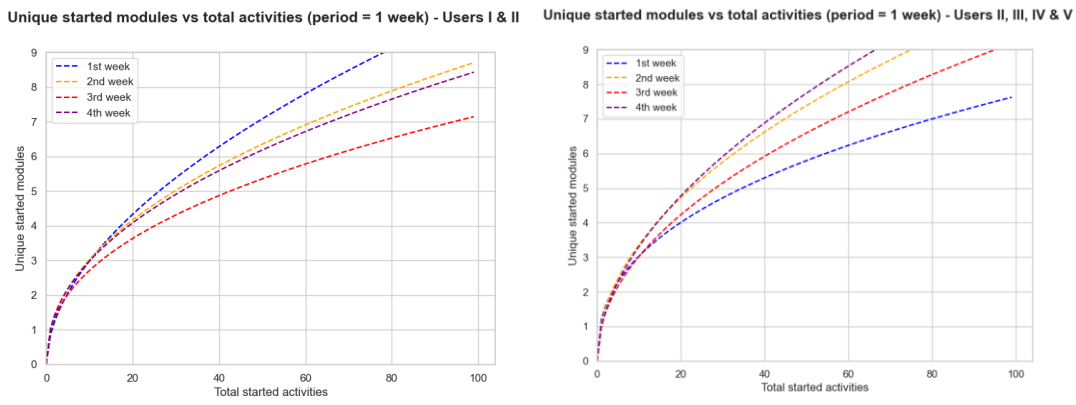
As for the evolution of the activities' entropy per types of users, the mean values in each time period are shown in the figure below, each with error bars indicating the SEM as a measurement of precision for the estimated means. One can see that the precision is well enough as error bars are not even perceivable, and that mean entropies do not considerably vary across time. In addition, almost all of the differences of entropy across the time periods have been in all cases proven to not be statistically significant (see Tables Table SI 4 and Table SI 5).



(a) (b)

***Figure 6**: Activities mean entropy evolution per user types (per days in the app) over time periods, calculated by applying subsampling for correction. Mean entropies are represented by dots and linked by lines, with colors specific for the user types. Additionally, SEM error bars (black vertical segments) are included for each of the means, indicating for precision in each case. Vertical axes correspond to activities entropies, and horizontal axes represent each time point, in a way that in the first one (value 1) the entropies correspond to the mean entropies in the first time periods (1 or 2 weeks). **(a)** Time period of 1 week. **(b)** Time period of 2 weeks.*

4.2.2 Modules



(a) (b)

***Figure 7**: Power law fits for the unique started modules as a function of the total started activities, in an aggregated manner per time periods of 1 week. **(a)** Considering user types I and II. **(b)** Considering user types II, III, IV and V. Associated p-values for both in **Table SI 6** and **Table SI 7**.*

The plots in Figure 7 aim at allowing to look at the EE behaviors regarding modules by means of comparing the tendencies of how many different modules users are expected to test from week to week as a function of the total started ones. See Figure SI 6 for the corresponding scatter plots

with the power law fits and Figure SI 7 for the results with time periods of 2 weeks, where it can also be appreciated that the fits in this case have a more reduced accuracy (R^2 values around 0.5).

In case of user types I and II (Figure 7a), statistical differences are found between the first week and the 2nd and 4th, as well as between the 2nd and 3rd. It can be observed how the power law fit gets flattened from the first week (blue dashed line) to the second and third (yellow and purple), indicating that users transition from testing more different modules to less. The same happens from the second to the third week (yellow to red). As for users staying more than 28 days in the app (Figure 7b), only the first week is statistically significantly different to the other time periods, and it can be seen that such users tend to the contrary behavior: to increase the number of different tested modules with time.

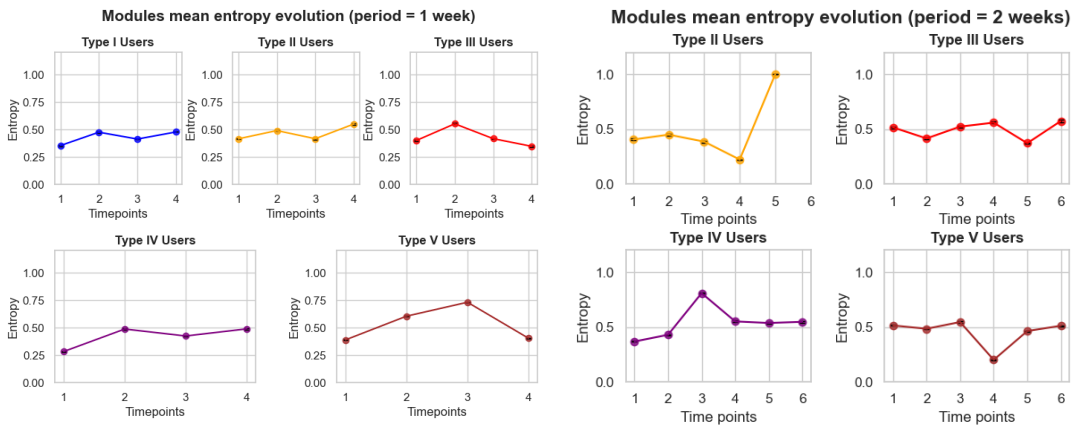


***Figure 8**: Modules mean entropy evolution per user types (per days in the app) over time periods, calculated by applying subsampling for correction. Mean entropies are represented by dots and linked by lines, with colors specific for the user types. Additionally, SEM error bars (black vertical segments) are included for each of the means, indicating for precision in each case. Vertical axes correspond to modules entropies, and horizontal axes represent each time point, in a way that in the first one (value 1) the entropies correspond to the mean entropies in the first time periods (1 or 2 weeks). **(a)** Time period of 1 week. **(b)** Time period of 2 weeks.*

As for the evolution of the modules' entropy, one can observe similar results as those for the activities' entropy: error bars are not perceivable, indicating precision, and the entropy mean values do not vary meaningfully or follow a pattern across time. Again, almost all the entropies differences across the time periods are not statistically significant (see Tables Table SI **8** and Table SI **9**).

4.3 Engagement analysis

The results of this analysis consist of entropies plots, now splitting the users into the engagement level groups. In such plots, the entropy evolution of each user types is plotted together in a same figure. The figures exposed here show the evolution from week to week, which are more interpretable than those obtained when using 2 weeks as time period (see Figures Figure SI 8 and Figure SI 10).

4.3.1 Activities

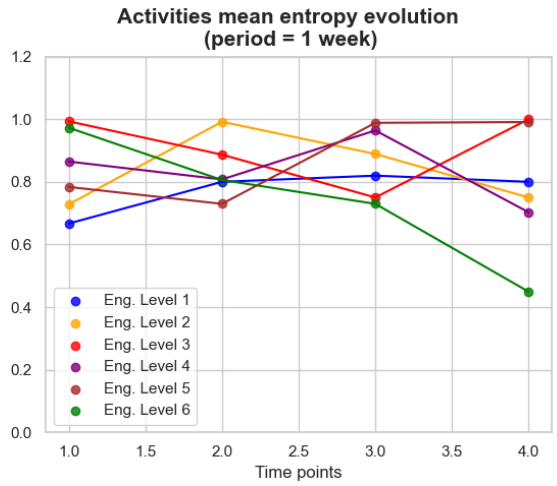


***Figure 9**: Unique plot containing, for each user type (per engagement level), the activities mean entropy evolution over time periods of 1 week, calculated by applying subsampling for correction. Colors represent the different user types, the vertical axis corresponds to activities entropies, and the horizontal axis represents each time point, in turn representing the weeks. Respectively in weeks 1 and 6, p-values with Kruskal Wallis are 0.6186 and 0.1592 (not statistically significant).*

In Figure 9 above it is seen how, in general, entropy is constant in time, without meaningful differences between the groups. Some remarkable details to note are that the least engaged users show the lowest mean entropy value in the first week, and the most engaged have the lowest value in week 4. However, Kruskal Wallis tests indicate no statistical significance between any of the groups in any of the two time points.

4.3.2 Modules

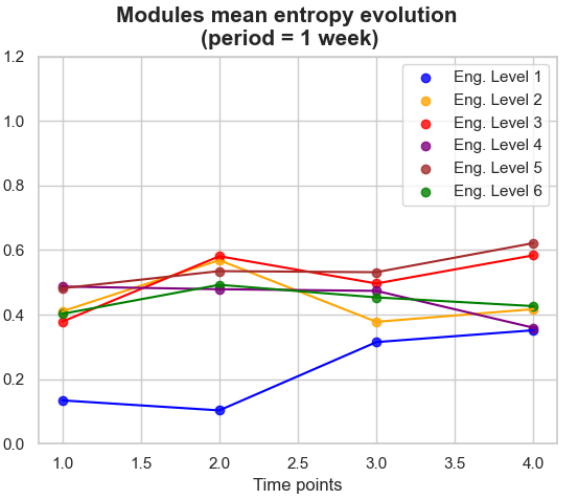


***Figure 10**: Unique plot containing, for each user type (per engagement level), the modules mean entropy evolution over time periods of 1 week, calculated by applying subsampling for correction. Colors represent the different user types, the vertical axis corresponds to modules entropies, and the horizontal axis represents each time point, in turn representing the weeks. In the first week, the calculated p-value with Kruskal Wallis is 0.037023 (showing statistical significance). Post-hoc Dunn's test p-values collected in **Table SI 10**.*

As for the modules' entropy, it is again generally observed to be constant in time. However, it can be seen that the users' group with lower engagement level present the lowest mean modules' entropy values in all time points. Specifically for the first week, Kruskal Wallis and Dunn's tests determine that the difference between group 1 and all the others is statistically significant.

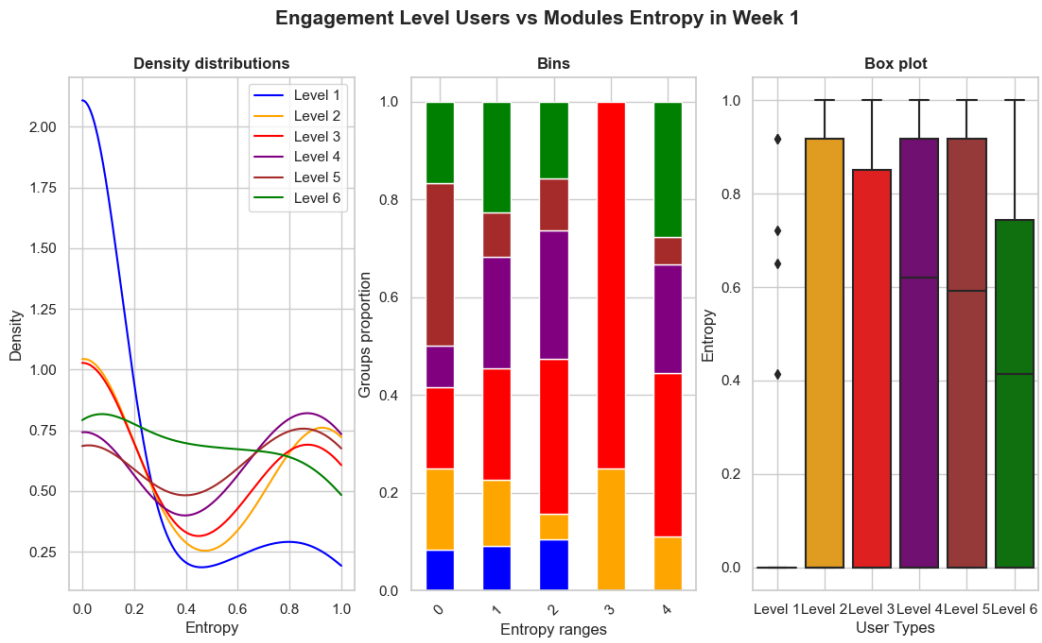


***Figure 11**: Set of three plots for getting a visual idea of whether modules' entropy in week 1 has predictive power to predict engagement. **(a)** Density distributions of modules' entropy in week 1 for each user group (see Figure SI 9 for probability histograms). **(b)** Proportion of groups with each specific range of modules' entropy. Ranges, in order, correspond to consecutive semi-open intervals with the values 0, 0.592, 0.811, 0.918, 0.995 and 1. **(c)** Box plot (for facilitating comparisons) showing the distribution of the entropy values across the different user groups.*

The figure above contains three plots that aim at trying to determine whether modules' entropy in the first week can be predictive on the engagement score. One can note that the distribution of such entropies especially differs from group 1 to the rest, which can be also clearly noted in the probability histograms in Figure SI 9. Even though the data is shown to be dispersed, in the sense that there are users of almost all types with entropies in a wide range of entropy values (Figure 11b and c), there is the exception of the least engaged users' type. Actually, in this latter case, the box plot indicates that users with an entropy different from zero are determined to be outliers.

This final scatter plot presents the engagement score versus the modules' entropy in the first week, with a fit to a polynomial of degree 1. The slope of the fit is positive, indicating that the higher the entropy, the higher the engagement score, specifically by a factor of 8.096 (slope).

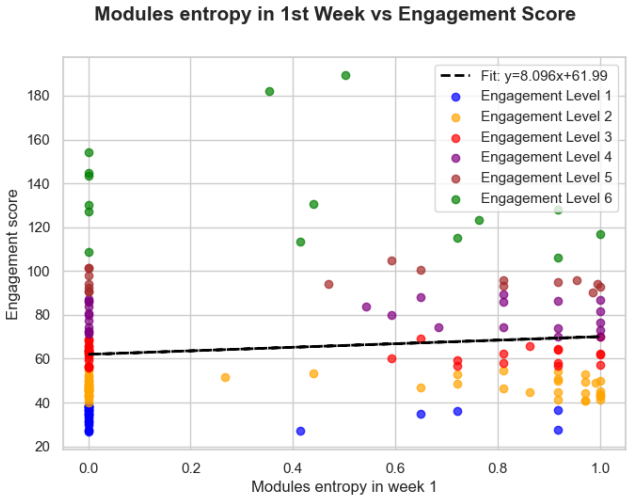


***Figure 12**: Scatter plot of engagement score (vertical axis) as a function of modules entropy (horizontal axis) in the 1st week, with a 1D polynomial fitted function.*

5. Discussion

User retention and engagement in Foundations has been observed to be deficient. Similar to the datum noted by Koa Health, only 682 out of the 919 total selected users (i.e., 74%) have been detected to not start activities after a month from their onboarding, which can be interpreted as stopping using the app. Because of this, the worth to analyze their behavior was indeed confirmed, with the aim of identifying trends in users that can be considered to have different levels of engagement. Relevant results of such analysis could potentially provide meaningful insights that helped taking strategies for improving the app's success among users. In this section, an interpretation of the results demonstrated to be statistically significant is done, accompanied by stating some of the limitations of the present thesis as well as related further work opportunities.

- **EDA**

The first analysis tackling metrics such as the total number of started and finished activities showed, in a statistically significant manner, a meaningful reduction in time of all metrics. This is remarkably seen from the first time period (either the first week or the first two weeks) to the others, indicating that users reduce their app usage at early stages of their stay. It is due to this fact that the results obtained throughout the full analysis when using 2 weeks as time period did not generally allow for visually extracting significant information and interpretations. Instead, it was needed to use one week as the time period, which has delimitated the analysis of the users' behaviors to their early stay stages, namely the first month (4 weeks). This approach is compatible and coherent with the available information in the dataset, in the sense that there is much more data of users starting activities in the first 4 weeks, simply because there is a higher number of them that remain this time in the app. In addition, taking into account their early renouncement and that the goal is to improve their engagement, it makes sense to study their behavior in the period of time they are more likely to indeed still be using the app. However, this makes up one of the major limitations of the work, as the EE dilemma is expected to appear once the decision-maker has gathered enough information to have been able to find valuable options that are potential for being exploited, rather than at the beginning of the process when multiple options are evaluated to gain knowledge about them, a pure exploration phase.

Another interesting information extracted was that none of the users estimated to remain in the app for over 260 days starts any activity before approximately a month and a half since his onboarding. Nevertheless, it has been observed that this fact is not a consequence of them not using the application, as they do generate events, which can be interpreted as carrying out other types of exploration rather than the assumed in this work. In further work, an analysis of which events do users that stay the longest time in the app generate before starting evaluating the options would be crucial.

- **Exploration-exploitation behaviors**

The interpretations extracted from the EE analysis only cover those plots of the first part, consisting of the power law fits to the scatter plots, as the entropy analysis both for the activities and for the modules showed no statistical significance. Because of this, entropy, which represented the metrics independent to the total number of started activities, was not reliable for being used to analyze users' behaviors. This way, a crucial point to mention for the evaluation of the scatter plots and power law fits results is that they are dependent on those total started activities. Further research should attempt to stumble upon a measurement unrelated to such biasing property. Those power law fits, which are the ones being compared for the identification of exploration and exploitation behaviors, are the result of fitting real users' data, dependent on how many activities each of them start. Moreover, in some of the cases, there is an excessive scattering of the data, which results in low-quality fits, especially for the modules analysis part. In addition, due to the limitation of users progressively leaving the app, as time goes by in the time periods, the number of data points gets reduced, and fits become less and less representative.

Consistently with what has been stated, statistical significance has been generally observed only between the first week and the others. The behavioral interpretations extracted from the plots differ between the two aggregated analyzed data. When studying aggregated data of users staying in the app over 28 days (i.e., user types II, III, IV and V), the general tendency seems to be the following: such users explore activities the most in the first week, experiencing a transition towards exploiting them the most on the second week, and then returning back to the initial behavior at weeks 3 and 4. As for the modules, they tend to start exploiting them and to show more and more exploration-inclined conducts as time goes on, progressively from the first to the third week. This way, users seem to start exploring activities and exploiting modules, and move towards the opposite behavior, that is, an exploitation of activities and an exploration of modules. However, when separately analyzing how do type I and II users (those from the types that stay fewer days) behave, the contrary conducts are identified. They are observed to explore activities the most on the second week, in which the fit actually indicates that all the activities they start are different. Regarding modules, they start mostly exploring and steadily tend towards their exploitation. Thus, users I and II in their first week do not explore *activities* the most, but indeed explore *modules* the most, tending to exploit them from week to week.

It may seem that a constraint of plotting aggregated data is that it does not allow for studying each case, when in fact, it is. However, the exposed graphs are already a proof that users' behavior is varied, such that there is not a consistent repeated EE conduct pattern in time with regard to neither activities nor modules. Actually, although not reported in this work, the behavioral tendencies of each of the user types were separately analyzed and observed to diverge. One first possible interpretation of the non-identification of such common EE behavior might be that users tend to explore and exploit activities and modules, but each of them doing so in different time periods, in kind of a totally user-dependent manner. Another explanation could be that, as previously mentioned, because of inspecting their first month of the app usage, users are simply exploring the content, and that the observed transitions in time are not in reality linked to EE practices but instead different manners of exploring.

- **Engagement analysis**

Originally, this second part of the in-depth analysis had the purpose of linking identified EE behaviors to users' engagement. Due to the rejection of the hypothesis about the existence of a pattern of such conducts, the engagement analysis was somehow restricted. As mentioned before, what is most important is in fact correlate users' behavior at initial stages, as it is when they are most probable to still not abandon the app, rather than the evolution of their app usage. This way, the focus was put on trying to predict users' engagement by means of some indicator of their early behavior.

When using the entropy metrics, no statistical significances were found with regard to the variability of started activities. On the contrary, some remarkable findings were noticed in case of modules' entropy. It was observed that users with the lowest engagement level showed meaningfully lower mean modules' entropies in all four weeks, meaning that they do not explore modules as much as the other groups. Such users' entropies (from which the mean was calculated) were proven to be statistically significantly different to those of the rest of engagement level groups in the first week of the application usage. Further generated graphs to determine whether this measurement has predictive power to forecast users' engagement indicated that, even though there are users of all levels of engagement for virtually all entropy ranges, an exception is detected for the least engaged users: they are observed to show significantly lower entropies, displaying their distribution a clear shift towards the minimum entropy (that is, 0). Actually, only three users present an entropy value different from 0, who are detected as outliers, and none of them reaching the maximum entropy. This predictive potential is ultimately confirmed with the polynomial function fitting the engagement punctuation as a function of entropy in week 1. Because of its positive slope, the feature of modules' entropy in the first week can be affirmed to be predictive of engagement.

A limitation that is worth to be mentioned is related to the establishment of the groups of users per engagement level. The created scoring system was arbitrarily established by means of own criteria, although combining reasonable engagement metrics as are the total stay days and percentage of days entering the app. One constraint linked to this lies in the fact that due to the usage of the engagement score for rating long-term engagement, which shapes an extracted feature, the correlation established between modules' entropy and engagement does not allow to predict a variable directly measurable from users' interactions with the app, such as the total number of days they will remain in it. Additionally, multiple other engagement metrics could have been used, such as users' estimated spent time on the app or number of sessions, and compared between each other to identify the most potentially predictable. Other further work would be to try to identify alternative relevant features, apart from the one identified here, indicating users' recurrent practices that had powerful predictive potential over their level of engagement, in order to use them to construct a machine learning model by training and testing its predictive capabilities.

- **Implications and further limitations**

The obtained results make up a thorough analysis of the usage of Foundations from which one could potentially extract meaningful information for the app improvement. The detected particular feature noted to affect users' engagement and thus identified as predictive intrinsically implies the following three certainties. In the first week:

1) if a user starts activities only from a single same module (i.e., minimum entropy = 0), he/she is likely to have a poorer engagement
2) if a user tries as many modules as possible (i.e., maximum entropy = 1), he/she will not be considered to have the poorest engagement
3) the more modules a user assesses, the higher his/her engagement level will be

This constitutes valuable information that could be potentially used to drive business decisions. The main action that could be taken to improve users' engagement would consist in inciting them to explore modules to the fullest possible. Koa Health makes use of two types of recommendation systems: one random, and another based on reinforced learning. Such systems already implemented in Foundations could be used to bias users' taken search strategies in their decision-making process by biasing them towards the decision of doing activities from distinct modules. In principle, only by making sure that the activities they start imply a minimum variability of modules, it would be ensured for the user to not be presenting the lowest engagement possible. Not only that, but also the higher such variability of modules was, the higher the probability for them to potentially show higher compromise, either as regards to the percentage of days entering the app or the total number of stay days.

Apart from the already declared limitations, it should be noted that recommender systems may be already biasing users' behavior. One of the main encountered difficulties was considering all relevant application features. For instance, activities are not only grouped into modules but also into programmes, which gives rise to the possibility of users showing EE conducts at the level of those programs. The fact of being unaware of users' objectives configures part of another principal limitation: several assumptions have been needed to be done with regard to the decision-making process elements, being the main one that their aim is to find the modules or activities that work best for them. However, this might be unrealistic to some extent, especially knowing that there are distinct types of offered activities (e.g., consisting of audios, videos or reading or writing tasks), being some of them intrinsically more improbable to be repeated. Furthermore, users' preferences are likely to vary over time, for example by them getting bored of repeating one same activity or module that used to suit them in the past.

Finally, external limitations linked to the interaction with Koa Health should be mentioned. The collaboration was closed in late 2021, and as previously stated, we received two versions of the data: the first one in March and the second in mid-May. In addition, the person in direct contact with us from Koa Health left the company in March and any other person acquired his role in the project. Naturally, all these contextual constraints have significantly affected the project development process.

Many further work opportunities than the specified thus far can be derived from this project, as an infinite number of alternative approaches different to the one established in this thesis could be taken to tackle the same problem, even from the EE perspective too. Additionally, a large number of extra data science analysis studies could be performed to identify patterns that allowed to support business decisions aiming to improve the startup product.

6. Conclusion

By having successfully executed the planned data science process, the initial set objectives of the present thesis have been fulfilled. The analyses performed can be affirmed to have provided insights into the way users are making use of Foundations and have ultimately resulted in the main hypothesis rejection. Through an exploratory analysis, the decrease in time of a wide variety of metrics has confirmed the sharp reduction of the app usage, especially noted after the first week from the onboarding moment.

Exploration-exploitation conducts have not been recognized in users' practices, nor patterns of behavior have been identified. The main limitation of the thesis constitutes the fact that the EE dilemma is expected to appear once the decision-making process is minimally advanced, the bare minimum so that a first pure exploration phase of gaining knowledge finishes. Because of most users dropping the app after a month of usage, there is missing data from which to extract how do users behave at further stages of their stay, in which they would be definitely more prone to expose conducts related to the trade-off under investigation. Multiple alternative and supplementary analyses ensued from the one carried out in this work could be performed, under distinct assumptions or analysis strategies, to continue chasing the identification of EE dilemma in this context.

Despite that, a relevant finding has been obtained with regard to the correlation between users' behavior and their commitment to the app. The feature regarding the variability of different started modules in the first week of the application usage has been shown to have predictive potential over engagement. This main extracted result may provide valuable information applicable from the business perspective to the app recommender systems, encouraging users to explore modules in order to increase their likelihood of engaging more and more with Foundations. Future work may focus on the extraction of further additional features, such as the one determined here, to complete the data science cycle through the construction of a machine learning model that allows to predict engagement and to further refine and boost the functionality of the existing recommender systems.

Bibliography

[1] Balke, T., & Gilbert, N. (2014). How Do Agents Make Decisions? A Survey. *Journal of Artificial Societies and Social Simulation, 17*(4). https://doi.org/10.18564/jasss.2687

[2] Trewartha, R., & Newport, M. (1982). *Management* (Third ed.). Dallas and Business Publication.

[3] Moreno-Bote, R., Ramírez-Ruiz, J., Drugowitsch, J., & Hayden, B. Y. (2020). Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences, 117*(33), 19799–19808. https://doi.org/10.1073/pnas.2004929117

[4] The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. (2015). *PLOS ONE, 10*(3), e0119116. https://doi.org/10.1371/journal.pone.0119116

[5] Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

[6] Lattimore, T., & Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

[7] Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science*, 322–331. https://doi.org/10.1109/SFCS.1995.492488

[8] Deci, E. L., & Ryan, R. M. (2013). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer Publishing.

[9] Iyengar, S. S., & Lepper, M. R. (2000). When Choice is Demotivating: Can One Desire Too Much of a Good Thing? *Journal of Personality and Social Psychology, 79*(6), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

[10] Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281–299. https://doi.org/10.1037/h0032955

[11] Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

[12] Schaeffer, S. E., & Rodriguez Sanchez, S. V. (2020). Forecasting client retention — A machine-learning approach. *Journal of Retailing and Consumer Services, 52*, 101918. https://doi.org/10.1016/j.jretconser.2019.101918

[13] Foreseeing Employee Attritions using Diverse Data Mining Strategies. (2019). *International Journal of Recent Technology and Engineering, 8*(3), 620–626. https://doi.org/10.35940/ijrte.b2406.098319

[14] Kumar Dubey, A., Maheshwari, I., & Maheshwari, A. (2018). Predict Employee Retention Using Data Science. *International Journal of Electrical Electronics & Computer Science Engineering*, 67-72.

[15] Bitrián, P., Buil, I., & Catalán, S. (2021). Enhancing user engagement: The role of gamification in mobile apps. *Journal of Business Research, 132*, 170–185. https://doi.org/10.1016/j.jbusres.2021.04.028

[16] Pujol Torrens, M. (2022). Analysis of the Breadth-Depth Dilemma through the Users of Koa Foundations, a Mental Well-Being App [Bachelor's Thesis, Pompeu Fabra University].

[17] Fawaz Siddiqi, M. (2021, January 4). *Following the data science methodology*. IBM developer. Retrieved February 16, 2022, from https://developer.ibm.com/blogs/following-the-data-science-methodology/

[18] Munnangi, J. (2021, March 10). *Data Science project life cycle*. Medium, Co-Learning Lounge. Retrieved February 27, 2022, from https://medium.com/co-learning-lounge/complete-data-science-project-life-cycle-9eae6e4ed4c9

[19] Agarwal, S. (2018, February 9). *Understanding the Data Science Lifecycle*. Retrieved February 27, 2022, from https://www.sudeep.co/data-science/2018/02/09/Understanding-the-Data-Science-Lifecycle.html

[20] The pandas development team. (2020). pandas-dev/pandas: Pandas 1.0.3 (v1.0.3). *Zenodo*. https://doi.org/10.5281/zenodo.371523

[21] Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical *Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[22] Campbell, L. L. (1966). Exponential entropy as a measure of extent of a distribution. *Probability Theory and Related Fields*, *5*(3), 217–225. https://doi.org/10.1007/bf00533058

[23] Carcassi, G., Aidala, C. A., & Barbour, J. (2021). Variability as a better characterization of Shannon entropy. *European Journal of Physics*, *42*(4), 045102. https://doi.org/10.1088/1361-6404/abe361

[24] Espinosa-Paredes, G. (2021). Nonlinear BWR dynamics with a fractional reduced order model. *Fractional-Order Models for Nuclear Reactor Analysis*, 247–295. https://doi.org/10.1016/b978-0-12-823665-9.00007-9

[25] Węglarczyk, S. (2018). Kernel density estimation and its application. *ITM Web of Conferences, 23*, 00037. https://doi.org/10.1051/itmconf/20182300037

[26] Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

[27] Harris, C. R., Millman, K. J., van der Walt, S. J. et al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[28] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272.

[29] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

[30] Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, *68*(6), 540. https://doi.org/10.4097/kjae.2015.68.6.540

[31] Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and Application of the Kruskal-Wallis Test. *Applied Mechanics and Materials*, *611*, 115–120. https://doi.org/10.4028/www.scientific.net/amm.611.115

[32] Mahajan, A. (2016). Post hoc tests in analysis of variance. *Indian Journal of Occupational and Environmental Medicine*, *20*(2), 121. https://doi.org/10.4103/0019-5278.197552

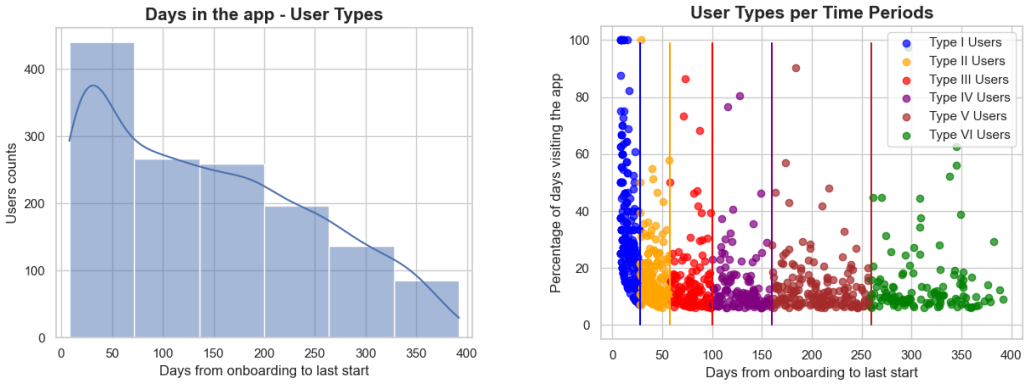
[33] Pedregosa *et al.* (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, pp. 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[34] Jaykaran. (2010). «Mean ± SEM» or «Mean (SD)»? *Indian Journal of Pharmacology*, *42*(5), 329. https://doi.org/10.4103/0253-7613.70402

Supporting information

1. Exploratory data analysis

1.1. User groups



(a) (b)

***Figure SI 1**: Visual representations used for the determination of the user groups per total days in the app, defined as the number of days from onboarding to the last time of starting an activity. **(a)** Histogram with counts of users per number of days in the app (of all users). **(b)** Scatter plot of percentage of days visiting the app as a function of the days in the app (for the 919 selected users), with vertical lines in the determined thresholds to separate users into the 6 groups. Colors of the data points indicate the assigned user type per days in the app for each user.*

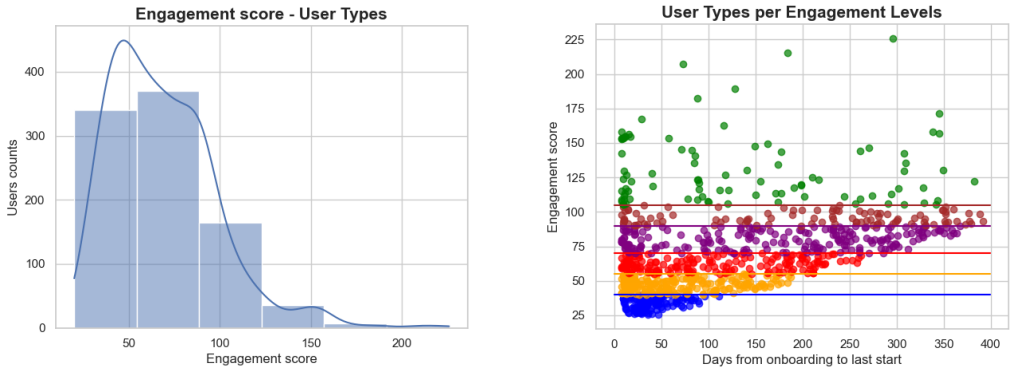


***Figure SI 2**: Visual representations used for the determination of the user groups per engagement level, determined by means of their associated engagement score. **(a)** Histogram with counts of users per engagement score. **(b)** Scatter plot of engagement score as a function of the days in the app, with horizontal lines in the determined thresholds to separate users into the 6 groups. Colors of the data points indicate the assigned user type per engagement level for each user.*

1.2. Metrics overall view

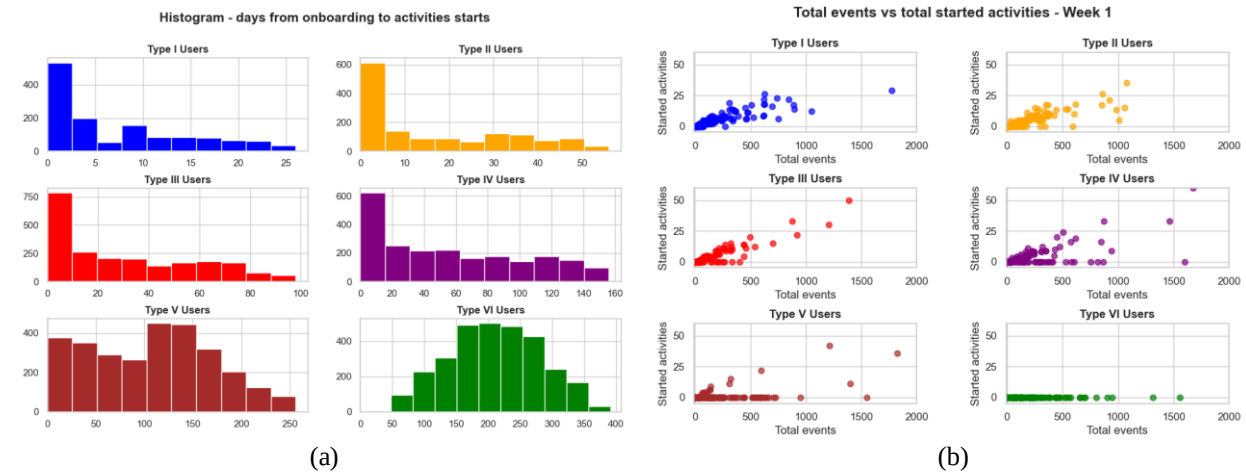
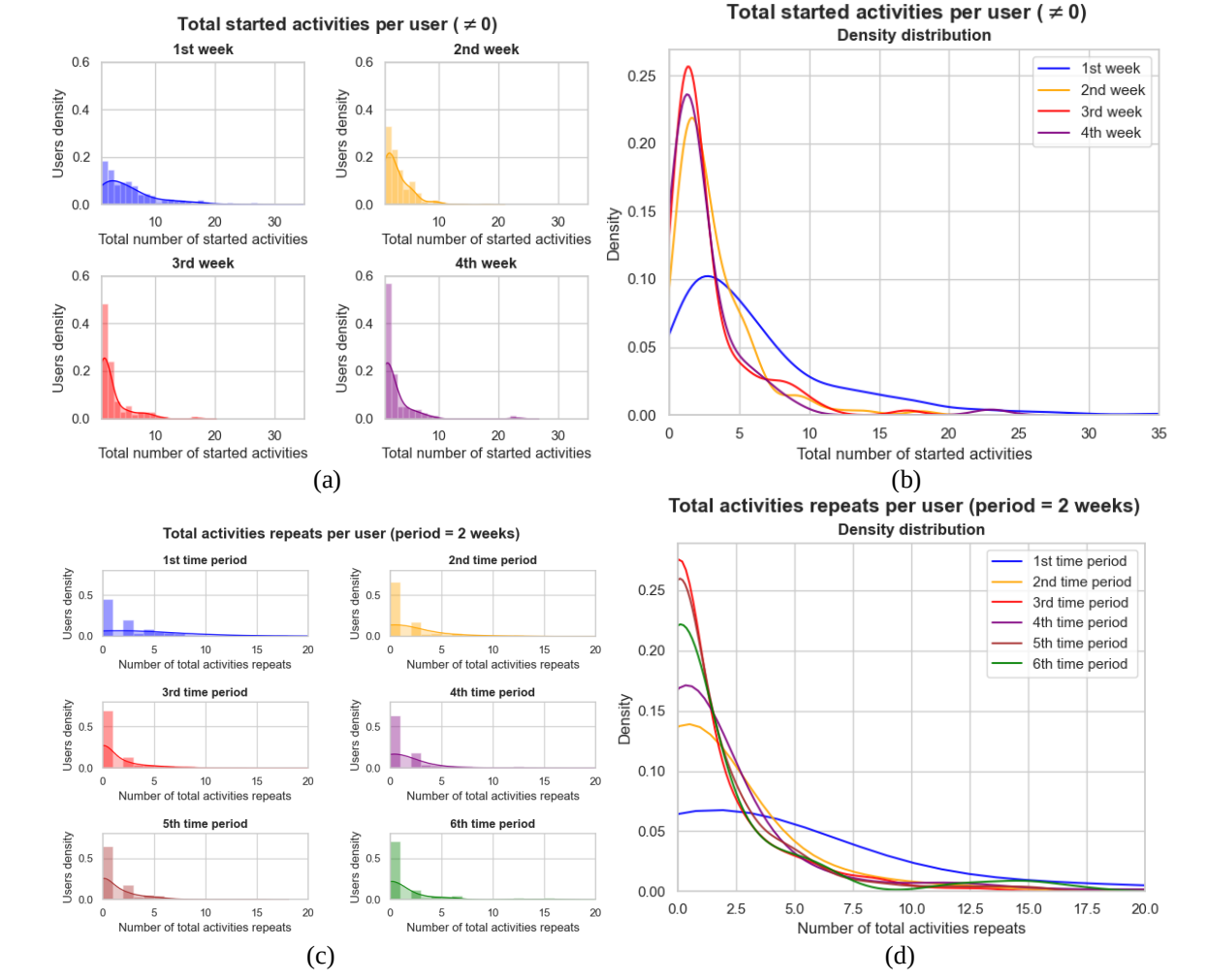


***Figure SI 3**: (a) Histograms per user types indicating the difference in number of days from the onboarding to those days in which they start an activity. (b) Scatter plots per user types for the number of started activities as a function of the total number of generated events, in the first week.*



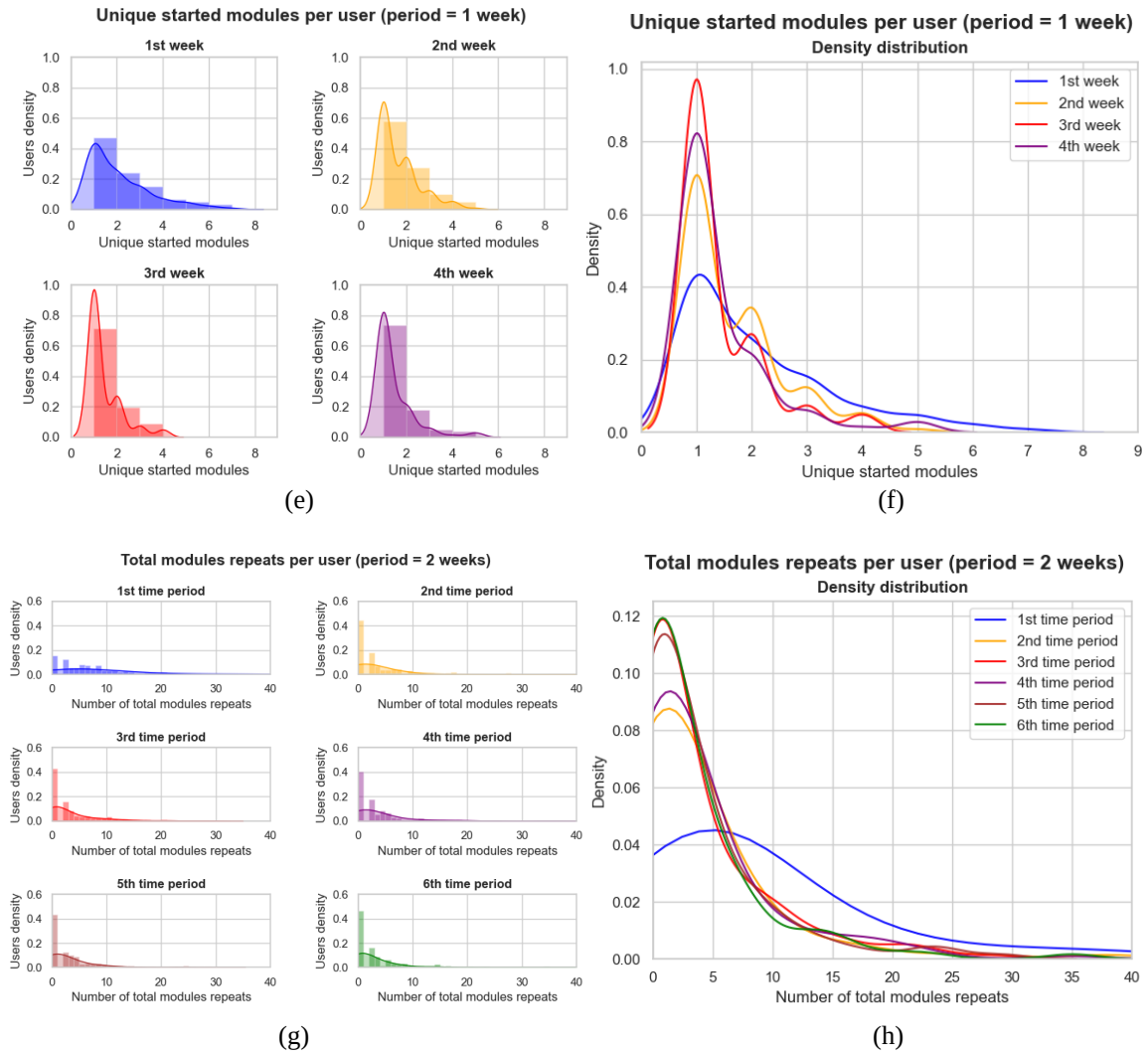


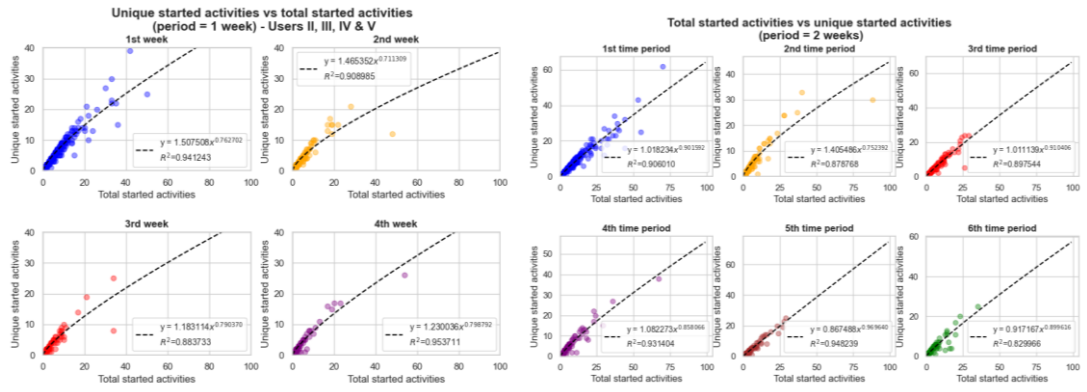
***Figure SI 4**: Collection of the most meaningful results for the overall view of the data part of the EDA, considering data from the total 919 selected users. For each metrics, histograms are plotted on the left and density distributions on the right. **(a)** and **(b)** Total number of started activities per week neglecting zeros for visualization. **(c)** and **(d)** Total number of activities repeats, i.e., sum of each number of starts of a same activity higher than 1, per time periods of 2 weeks. **(e)** and **(f)** Number of unique (different) started modules per week. **(g)** and **(h)** Total number of modules repeats, i.e., sum of each number of starts of a same module higher than 1.*

2. Exploration-exploitation analysis

2.1. Activities

WEEKS	1	2	3	4
1	1.0000	$1.6796e^{-12}$*	$5.524e^{-8}$*	$1.0068e^{-12}$*
2		1.0000	0.00117*	0.4320
3			1.0000	0.02571*
4				1.0000

***Table SI 1**: p-values, from the pairwise Wilcoxon tests, comparing aggregated data (of users I and II) of the number of unique started activities as a function of the total started activities, per week (**Figure 4**b). *comparisons showing statistical significance (i.e., p-value≤ α=0.05)*



(a) (b)

Figure SI 5*: Scatter plots showing, in an aggregated manner for user types II, III, IV and V and per time periods, the number of different started activities (vertical axis) as a function of the number of total started activities (horizontal axis), each together with their associated fitted power law function (black dashed lines). Unique plots with the power law fits in Figure 5.* ***(a)*** *Time period of 1 week.* ***(b)*** *Time period of 2 weeks.*

WEEKS	1	2	3	4
1	1.0000	$8.4110e^{-18}$*	$6.8210e^{-19}$*	$6.2413e^{-19}$
2		1.0000	0.13609	0.23317
3			1.0000	0.7858
4				1.0000

Table SI 2: *p-values, from the pairwise Wilcoxon tests, comparing aggregated data (of users II, III, IV and V) of the number of unique started activities as a function of the total started activities, per week (Figure 5a and Figure SI 5a). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

TIME PERIODS	1	2	3	4	5	6
1	1.0000	$1.5226e^{-07}$*	0.0148*	0.4907	0.9249	0.0538
2		1.0000	0.005219*	0.06673	0.03441*	0.001033*
3			1.0000	0.006204*	0.0069734*	$1.069e^{-05}$*
4				1.0000	0.008546*	$1.428e^{-04}$*
5					1.0000	$7.199e^{-04}$*
6						1.0000

Table SI 3*: p-values, from the pairwise Wilcoxon tests, comparing aggregated data (of users II, III, IV and V) of the number of unique started activities as a function of the total started activities, per 2 weeks (Figure 5b and Figure SI 5b). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

- **Entropy**

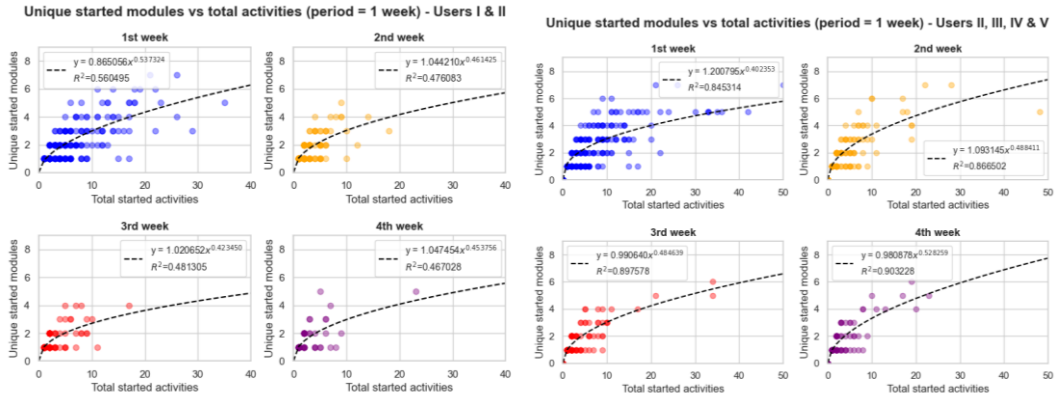
USER GROUP	WEEKS 1&2	WEEKS 2&3	WEEKS 3&4
TYPE I	0.6311	0.0011*	0.7033
TYPE II	0.1510	0.5155	0.3796
TYPE III	0.3662	0.0656	0.03389
TYPE IV	0.1486	0.2439	0.4698
TYPE V	0.5637	0.3173	0.6547

Table SI 4: *p-values, from the pairwise Wilcoxon tests, comparing users' activities entropies in the consecutive time periods of 1 week (**Figure 6**a). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

USER GROUP	TP 1&2	TP 2&3	TP 3&4	TP 4&5	TP 5&6
TYPE II	0.0973	0.7389	0.0137*	0.00389*	0.1573
TYPE III	0.01826*	0.7172	0.7502	0.3008	0.7354
TYPE IV	0.1136	0.5637	0.6148	0.7055	0.3490
TYPE V	0.1573	0.5809	0.4497	0.7389	0.8907

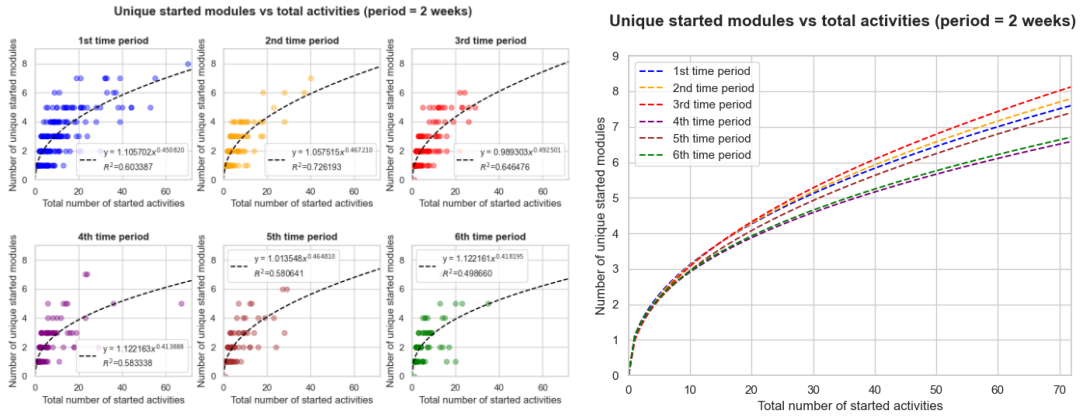
Table SI 5*: p-values, from the pairwise Wilcoxon tests, comparing users' activities entropies in the consecutive time periods of 2 weeks (**Figure 6**b). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

2.2. Modules



(a) (b)

Figure SI 6*: Scatter plots showing, in an aggregated manner per time periods of 1 week, the number of different started modules (vertical axis) as a function of the number of total started activities (horizontal axis), each together with their associated fitted power law function (black dashed lines). Unique plots with the power law fits in Figure 7.* ***(a)*** *Considered user types I and II.* ***(b)*** *Considered user types II, III, IV and V.*



(a) (b)

***Figure SI 7**: **(a)** Scatter plots showing, in an aggregated manner for user types II, III, IV and V and per time periods of 2 weeks, the number of different started modules (vertical axis) as a function of the number of total started activities (horizontal axis), each together with their associated fitted power law function (black dashed lines). **(b)** Power law fits for each time period of 2 weeks.*

WEEKS	1	2	3	4
1	1.0000	0.03498*	0.4238	0.002355*
2		1.0000	0.003220*	0.37432
3			1.0000	0.08402
4				1.0000

***Table SI 6**: p-values, from the pairwise Wilcoxon tests, comparing aggregated data (of users I and II) of the number of unique started modules as a function of the total started activities, per week (Figure 7a and Figure SI 6a). *comparisons showing statistical significance (i.e., p-value≤ α=0.05)*

WEEKS	1	2	3	4
1	1.0000	$9.9993e^{-04}$*	$5.17099e^{-05}$*	$1.62900e^{-03}$*
2		1.0000	0.36808	0.8327
3			1.0000	0.4587
4				1.0000

***Table SI 7:** p-values, from the pairwise Wilcoxon tests, comparing aggregated data (of users II, III, IV and V) of the number of unique started modules as a function of the total started activities, per week (Figure 7b and Figure SI 6b). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

- **Entropy**

USER GROUP	WEEKS 1&2	WEEKS 2&3	WEEKS 3&4
TYPE I	0.8154	$4.5701e^{-05}$*	0.5002
TYPE II	0.9340	0.6687	0.9280
TYPE III	0.1505	0.8884	0.1089
TYPE IV	0.0186*	0.7172	0.4784
TYPE V	0.2932	0.2850	0.7505

Table SI 8: *p-values, from the pairwise Wilcoxon tests, comparing users' modules entropies in the consecutive time periods of 1 week (Figure 8a). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

USER GROUP	TP 1&2	TP 2&3	TP 3&4	TP 4&5	TP 5&6
TYPE II	0.3458	0.9082	0.0037*	0.0422*	0.1573
TYPE III	0.6859	0.2978	0.6995	0.4380	0.9455
TYPE IV	0.4309	0.0560	0.9392	0.9322	0.4977
TYPE V	0.8927	0.9052	0.0750	0.2924	0.4313

Table SI 9*: p-values, from the pairwise Wilcoxon tests, comparing users' modules entropies in the consecutive time periods of 2 weeks (Figure 8b). *comparisons showing statistical significance (i.e., p-value≤ α=0.05).*

3. Engagement analysis

3.1. Activities

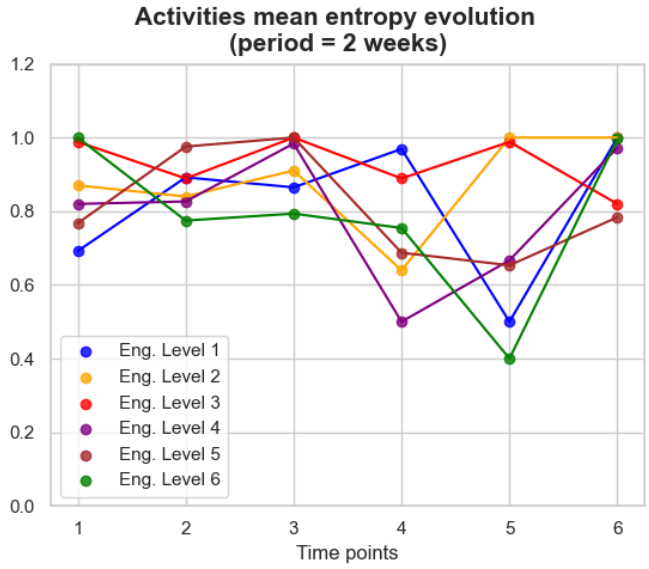


Figure SI 8*: Unique plot containing, for each user type (per engagement level), the activities mean entropy evolution over time periods of 2 weeks, calculated by applying subsampling for correction. Colors represent the different user types, the vertical axis corresponds to activities entropies, and the horizontal axis represents each time point, in turn representing the time periods.*

3.2. Modules

GROUPS	1	2	3	4	5	6
1	1.0000	0.005133*	0.030525*	0.002746*	0.010041*	0.045675*
2		1.0000	0.635770	0.500772	0.627191	0.950512
3			1.0000	0.308570	0.420194	0.781643
4				1.0000	0.933998	0.580145
5					1.0000	0.664731
6						1.0000

***Table SI 10**: p-values, from the Dunn's test, comparing modules mean entropies of each engagement user type groups in week 1. *comparisons showing statistical significance (i.e., p-value≤ α=0.05)*

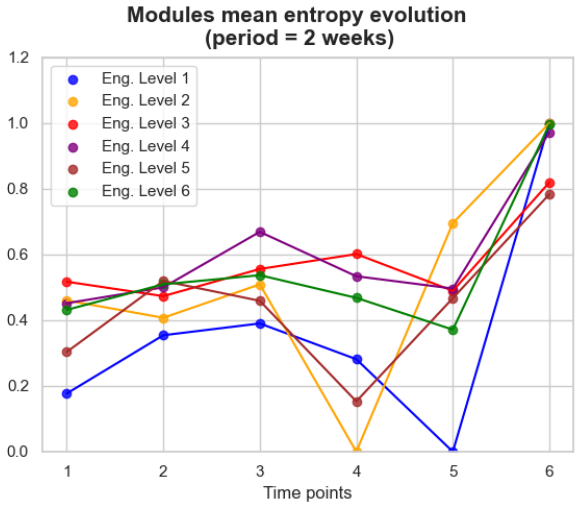


***Figure SI 10**: Unique plot containing, for each user type (per engagement level), the modules mean entropy evolution over time periods of 2 weeks, calculated by applying subsampling for correction. Colors represent the different user types, the vertical axis corresponds to modules entropies, and the horizontal axis represents each time point, in turn representing the weeks.*

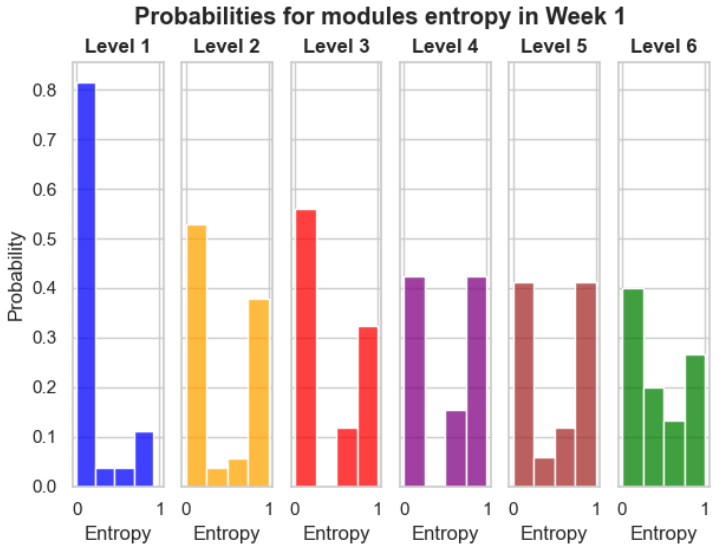


***Figure SI 9**: Probability histograms of modules' entropy in the 1st week for each of the engagement level user types. It can be observed how engagement level 1 users show the highest probability for lower entropy values.*