August 16, 2019

Data Science
Editorial board

Dear Editors and Reviewers,

We would like to thank you for the useful comments that permitted us to greatly improve the manuscript. In the current version of the work, we have seriously taken into account each comment and have improved the paper accordingly. Below, we explain how we addressed the editor's and each reviewers' comments. For easing the reviewing work, the major changes are marked with blue text in the revised version of the manuscript.

Meta-review " We encourage you to revise the manuscript, based on the 2 reviews and 1 comment made."

> **Our answer:** In the following, we discuss in details how we addressed the concerns raised by the Reviewers.

Review 1

R1.1 " Summary of paper in a few sentences:

The paper tests four models (Cox Proportional Hazard Model, Negative Binomial Model, Random Forest Model, XRBoost Model) to predict mobile phone applications user engagement. The goal is to provide evidence that it is possible to predict user engagement fairly accurately, though the wider goal is to facilitate intervention to re-engage users or increase user engagement. The models are tested using data from user engagement with a waste recycling app. The paper shows that random forest and XGBoost models are best suited to make accurate predictions in terms of identifying engaged and disengaged users. The negative binomial model is the least accurate model in terms of predicting the number of user activities before disengagement. "

> **Our answer:** The Reviewer is right. Our aim is indeed to provide evidence of the possibility of predicting User Engagement. We test that hypothesis by means of a recycling app dataset. The Reviewer's remarks concerning the performance of the models are also accurate.

R1.2 " Reasons to accept:

The paper provides evidence that user engagement can be predicted, albeit I would like to have a more explicit discussion to what extent this evidence has been missing so far, given this seems to be the main novelty of the paper."

> **Our answer:** To address this issue we have added this to the introduction:
>
> > *"Despite there being a good understanding of what is UE in different domains and which factors contribute to it, there seems to be a lack of literature on whether it is possible to predict UE in mobile apps and how different methods perform."*
>
> That motivates and contextualizes the gap this paper is trying to bridge. In addition to that, we point out to our background discussion (2.1), where we highlight that even defining UE in our context (and generally speaking too) already poses a real challenge. We cite a significant number of references in that section and provide different ways to

measure/calculate it. We aim at shedding light on the fact that there is not a standard way of dealing with the choice of methodology, hence we provide results using different approaches, namely counting (NB) and recency (CPH, RF, XGBoost).

R1.3 " The paper deploys and compares four different models to make predictions on user engagement, each useful in its own right and it is indeed very insightful to see how these models can be used with these type of data. The authors do admit that the comparison across these four models is not entirely justified, given their different natures, however, this need further acknowledgement in the discussion section. In fact, only RF and XGBoost models are really reasonably comparable, since both aim at the classifying users into engaged/disengaged. The other two models have very different predictive objectives and setups and hence comparing them to each other and to the two classification models is highly problematic in my eyes."

Our answer: We agree with the Reviewer. We have addressed that point in our answer to R1.6.

R1.4 " The authors show the value of clustering users prior to the modelling to obtain better prediction results. However, I would like to know what variables were used for clustering and what the clusters actually mean."

Our answer:
We use all 122 variables available in our hierarchical clustering algorithm. The idea is to be able to represent different user characteristics such as geographical, app usage or point collection. We have added the following paragraph to the manuscript to address your remark:

> *As the last part for the configuration of our clustering, we choose which variables we consider to be used for clustering. The variables we pick determine what our clusters represent. As an initial set of variables for our clustering algorithm, we choose all 122 variables mentioned above. In this context, our clusters represent different characteristics of the users and their behaviour, ranging from regional data to frequency of use and point collection. Users in the same cluster are thus expected to be more similar when it comes to app behaviour and geographical location compared to those in other clusters. Hence, these clusters capture useful information for our different user engagement models to use in their predictions.*

R1.5 " Reasons to reject:

I find the framing of the paper problematic in term of the wider goal of this research. Why should the users be prompted to use an app, that they probably consider (temporary) irrelevant, why should we try to make people spend even more time sticking to their phones? I find these explicitly stated goals highly problematic. It is quite striking that when the authors list all kind of reasons why people choose to disengage with an app (p.2—), the most obvious reason, that the app has lost its relevance (at least temporary), is not even listed. Also, I see how this investigation is of interest to the industry, but, this is supposed to be an academic paper and I would like to see how this is relevant for science. The authors write on p.2 "...provide a framework for modeling and predicting UE, which can be further extended or used in other scientific studies", this needs to be significantly expanded. "

Our answer: To address the concerns mentioned by the Reviewer, we added the following paragraph to the introduction. Here, we aim at extending the applicability of UE from a purely industry-based scenario to the scientific community.

"A high disengagement rate is obviously non desirable to app developers, whose success depends on the usage of their app. Furthermore, it is also a problem for researchers and other professionals who use apps to provide services aimed at improving the user's quality of life. Let's look at the e-Health domain as an example. Within e-Health, apps are used as a tool to help users overcome their illness (physical and/or mental) and improve their quality of life. However, for the app to succeed, it must be regularly utilised by the user. Hence, as a crucial quality, it must be engaging."

In addition to that, in the background section, we use several papers to motivate our study (from an academic point of view). The definition of engagement plays a crucial role [13], as well as defining its life-cycle [14,15,16]. Last, [17] (and to some extent [18]) discuss the temporal evolution of engagement and applications to different areas (health and e-learning). We believe our paper contributes to the scientific community by showing that it is possible to predict UE in mobile apps with a good degree of accuracy. In addition, we shed light on 4 different methodologies of how to achieve that.

R1.6 " I am not convinced the four models should be compared at all (see comments above). I think the authors should rather treat the models in their own rights, given they serve different modelling/prediction purposes."

Our answer: We agree. Hence, we focus the comparison mostly between RF and XGBoost (as also pointed out in the R1.3). Given those two are both tree-based machine learning models, we can compare them to a larger degree. We do not intend to diminish the importance of binomial models and/or the CPH approach. As you assertively point out, the binomial model has even a different target variable (counts) if compared to the ML approaches (days to disengagement). We are interested, however, in understanding which type of approach is more suited to model this problem (i.e. counting actions until disengagement, or counting days until disengagement). One of our conclusions is that days to disengagement seem a more actionable metric from a re-engagement perspective. The manuscript reads:

"...We highlight that a direct comparison between numerical models is not always possible due to their different natures - classification and regression. Thus, we aim to characterize and evaluate them mostly individually. When possible, we try to place our results in a broader perspective."

"...Concerning RQ_2, we applied the four models on the dataset and analysed the results obtained, mainly via the use of ROC curves. All models performed well, in their own right, with Cox proportional hazards, random forest and the boosted-tree models resulting in similar performance when predicting user engagement."

"... CPH, RF and XGBoost models result in similar values of accuracy. Their AUC values are similar, ranging roughly from 0.8 to 0.9. Our fourth model, the NB model, resulted in an AUC of 0.67. It is important to re-iterate that this AUC values should be taken as individual measures of performance and not used to compare models, as the manner of predicting and even the element of prediction is different according to the algorithm used."

R1.7 " The clustering needs further explanation and interpretation (see comments above)."

Our answer: We have elaborated our clustering explanation in our answer to R1.4.

R1.8 " Further comments:

Please use gender-neutral (possessive) pronouns, e.g. on page 2 instead of "...better suited for his own mobile app", write "...better suited for their own mobile app" or on page 5 instead of "(2) the time of her last event within the app" (which by the way sounds a bit awkward anyway), write "(2) the time of their last event within the app" (as noted, you may want to rephrase the entire statement). "

> **Our answer:** Thanks for pointing that out. We agree with the suggestion. Hence, all of the possessive pronouns have been changed to be gender neutral: "...better suited for their own mobile app"
> "the time of the user's last event within the app,"

R1.9 " Figure 6, what you claim to be blue (predicted events) appears as black (at least on my screen). "

> **Our answer:** We thank the reviewer for bringing this to our attention. The caption to the figure has been changed to say black instead of blue.

Review 2

Summary of paper in a few sentences: " This submission presents a framework to model and to predict user engagement with mobile applications. The framework is evaluated by using a data set of app usage of one particular app focusing on waste recycling. With the successful evaluation, the authors aim to provide evidence that it is possible to predict when users of a mobile application will get disengaged. "

R2.1 " Reasons to accept:

The focused topic of modeling and predicting user engagement for mobile applications is timely and relevant for the research communities in Data Science and Human-Computer Interaction. Overall, the presented approach seems to be novel and well-suited. Furthermore, also the results of the evaluation are promising. "

> **Our answer:** We thank the reviewer for the encouraging feedback. We have addressed your specific points below.

R2.2 " Reasons to reject:

I have strong doubts regarding the used data set and features. The used waste recycling app is described only briefly. The authors do not argue, why this is a common mobile application. I would recommend discussing this with consideration of the results presented by Müller et al. [1]. I would question that it is common for mobile applications that gamification aspects (here earning points) are directly connected to providing monetary benefits. Here, it is particularly interesting that the granted points can only be used at local shops. Thereby user's location becomes an obvious feature for disengagement. Additionally, using the zip code and the geolocation provides only redundant information. In general, the list of features and calculated variables is fuzzy. The authors claim to use 7 features but present only 6 in a list."

> **Our answer:** We have expanded the paper to accommodate a more comprehensive discussion on the dataset and features, as requested by the Reviewer, including the reference to Müller et al. The design of a possible web application should indeed be optimized to the fact that the user is either at the recycling location or at a shop redeeming the points. Most likely though, the choice of device will be a mobile phone, in this particular case.
>
> > *Extending the framework described in [1] for tablets, we argue that the app needs to be designed and optimized having in mind that the user is most likely on their mobile phone either redeeming points at a shop or collecting points at the recycle bin. That is fundamental to create an intuitive interface that facilitates these activities and promotes engagement.*
>
> Gamification is the common aspect and one of the future possible applications is to study how people redeem the points (immediately after a minimum or after some accumulation) creating personas and tailoring notifications to these profiles.
>
> > *Analyzing user behaviour to predict and prevent disengagement certainly poses a significant challenge, both from the methodological and analytical points of view. Due to the complexity of this task, we limited this study to characterizing and evaluating our methodology to predict UE. In a follow-up study, we will investigate how to ultimately influence user behaviour by increasing re-engagement rates and decreasing disengagement. Moreover, further research will touch upon studying the re-engagement process. Ultimately, our future objective is to determine the most appropriate interaction for each user at any given time, aiming to augment usage and prevent dis-*

engagement. Understanding the role gamification plays in mobile apps is also crucial. It can be done by further investigating how people redeem their points earned (e.g. immediately after achieving a minimum threshold or after some accumulation). That information helps in determining the type of notification that can be sent to each user.

Geo location and zip code are complementary given that some zip codes may have more than one bin. Hence, lat/lon pairs can provide further refining to zip code. We have also added that information to the manuscript.

Note that geographical position provides more detailed information than just zip-code, given that there may be more than one recycle bin in a given area.

Lastly, we have corrected the list of features to reflect the 6 features used (instead of 7).

R2.3 " Also, it reminds unclear how they combined the features to the 122 variables. "

Our answer: Following R2.2, we have expanded that section of the manuscript to detail how the variables are created.

We expand each of the 27,000 entries of the dataset to contain 122 unique variables in total. We achieve that by first generating combinations of these variables, e.g. number of days since the first event during weekdays or time of the user's last event within the app during a weekday/weekend. We then proceed to calculate the following statistics (max/min/mean/med/sum/sd) for all of the variables. That allows for more feature creation, e.g. standard deviation of the number of days since the first event during weekdays. We calculate the most simple statistics such as mean of the current point balance or minimum number of days since last event, but also combinations of variables with statistics - such as median of the minutes since last event per user in a certain zip code, or the standard deviation of the number of days since the first event during weekdays.

R2.4 " The authors do not describe if the application triggered any notifications. However, Sahami Shirazi et al. describe notifications as an essential element for engaging with mobile applications [2]. Hence, I wonder why the authors did not use the number of notifications or the reaction to notifications as a feature. To be able to understand user engagement or disengagement with the waste recycling app, it would be helpful, if the authors would also publish the application or provide at least a reference to the application."

Our answer: At this moment, there is no push notification implemented - even though we are very much aware that the number of notifications and the time they are sent are critical features to be taken into account. To avoid any ambiguity, we have added that explicitly to the manuscript. We use this bibliographical reference to strengthen the claim that notifications can both maintain engagement and trigger re-engagement.

Moreover, further research will touch upon studying the re-engagement process. We then intend to use push notification information - extending on the work of [2] - to ultimately determine the most appropriate interaction for each user at any given time, aiming to augment usage (maintain engagement) and prevent disengagement. Understanding the role gamification plays in mobile apps is also crucial. It can be done by further investigating how people redeem their points earned (e.g. immediately after achieving a minimum threshold or after some accumulation). That information helps in determining the type of notification that can be sent to each user.

R2.5 " While the authors motivate their work in the introduction very general, also looking on specific application domains such as health (reference [9] in the submission), the authors discuss the limitation of the used data set only briefly at the end of the paper. "

> **Our answer:** We have further elaborated the discussion on the data set description to accommodate for that remark. You are right, and we do not want to claim general applicability of this methodology yet, as we recognize that more research is needed for such a strong claim. Here, we really want to convey the message that it is possible to accurately predict UE through a reusable framework.
>
> "...note the modelling results - especially the quantitative component - discussed here remain specific for this dataset. Hence, it should not be directly transferred to other application domains. Instead, the main contribution of this paper lies on the fact that we show, by means of different types of algorithms, that it is possible to accurately predict user engagement as well as a reusable framework that can be used to better understand UE in mobile apps."

R2.6 " Further comments:

As described the used features and the data set look more specific than general to me. Hence, the submission would be more substantial if the authors would make less general claims and focus particularly on comparable mobile applications. Furthermore, publishing not only the data set but also the application or providing a reference to the application would improve the validity. "

> **Our answer:** We agree with the Reviewer but, unfortunately, we are not allowed to share further details about the application nor the dataset. We have already used in the manuscript and shared with the journal everything we are allowed to. In the repo, we are sharing the plots as well as the tables used to make them allowing for reproducible results. Concerning the more specific claims, we believe our answer to R2.5 covers that particular point.

Comment

This paper studies the user engagement in mobile apps. This is an interesting problem with important practical implications. The paper investigates the predictability of when mobile app users get disengaged with apps and shows that it can achieves the engagement prediction with a good level of accuracy. It applies different prediction models and also show clustering further facilitate the prediction. The paper is interesting, but there are some limits.

C1 " First, it only shows the predictability while does not show much details about the prediction itself, namely, what features would lead to the prediction. As a result, the practical implication would be limited. Also the features used in the prediction might not provide useful indication of engagement management without further investigation, such as significance test, etc. "

> **Our answer:** We point out to Table 1, where we discuss the predictors themselves. There, we observe that Groups and number of actions are very important to the RF model to make predictions. Followed by location and time descriptors. Note that the same order follows for the boosted-trees algorithm (in terms of gain). In both cases, we notice that the clustering algorithm plays a fundamental role in augmenting the accuracy of the models. That is a very actionable conclusion to be drawn from this study. In addition, the number of actions is also important indicating that users need to remain relatively active over time to prevent disengagement. Further recommendations can be push notifications to keep the users engaged in the short term as well as redeeming points constantly to keep the gamification aspect active for as many users as possible. To account for your remark, we highlight these points below:
>
> *To further understand which processes/features determine the behaviour of this model, in Table 1 we show the mean decrease in accuracy (MDA) for some of the predictors. The MDA is calculated by permuting the values of each predictor and then measuring by how much the predictive accuracy decreases.*
>
> *In our case, removing groups, number of actions, longitude, or weekday, from the predictors list would decrease the accuracy of this model by over 30%.*

C2 " Second, the prediction just applies some standard models without much technical novelties (also given the practical implication can be limited given the current status of the paper (ie, lacking of details about the prediction model); it would be helpful also give more details of technical barrier of the problem and solutions. "

> **Our answer:** By design, the study was set out to be an 'exploratory' study, in which we:
>
> > "provided evidence that predicting when users of mobile apps get disengaged is possible with a good level of accuracy."
>
> Hence, using more standard predictive models would provide a preferred basis on which future work could expand upon. We believe that at this stage, the focus should be on proving evidence, we are able to predict UE with reasonable accuracy in this particular case. We do agree with the remark that at a later stage, newer/modified models can be applied to derive different insights. We are also interested in experimenting with push notifications, i.e. how does the content and timing of such messages augments/diminishes engagement. We have the clustering results to already provide a solid basis on which groups to focus first and later expand on that. However, that remains out of the scope of this paper.

C3 " Third, more features would be helpful, in particular some features can explain use engagement such as version updates, similar apps in the market. App usage feature might just related to what

to predict in this paper. ”

> **Our answer:** We agree with the Reviewer that it would be interesting to investigate how the use of more/different features can influence the models and their performance. However, for this work, other features were not available as the app did not collect them. Here, we used all features were made available to us. We have experienced with creating extra features (122 in total) with the variables available to us as described in 2.2, however, higher-order statistics or different combinations of the features described in 2.2 would not improve predictability nor accuracy. Here, we explicitly decided to limit the scope of this study to show UE can be predicted with a reasonable degree of accuracy. Future collaborations can aim at determining/investigating what groups of features can help explain UE even further.

Best regards,

Eduardo Barbaro, Eoin Martino Grua,
Ivano Malavolta, Mirjana Stercevic,
Esther Weusthof, Jeroen van den
Hoven

References

[1] Hendrik Müller, Jennifer Gove, and John Webb. Understanding tablet use. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. ACM Press, 2012.

[2] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. Large-scale assessment of mobile notifications. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM Press, 2014.