Classifying Crowdsourcing Platform Users' Engagement using Machine Learning and XAI

Sana Hassan Imam
Institute of Computer Science, Faculty of Business Studies and Economics, University of Bremen
Germany
imam@uni-bremen.de

Christopher A. Metz
Institute of Computer Science, University of Bremen
Germany
cmetz@uni-bremen.de

Lars Hornuf
Faculty of Business and Economics, Technische Universität Dresden, Germany
lars.hornuf@tu-dresden.de

Rolf Drechsler
Institute of Computer Science, University of Bremen, Cyber-Physical Systems, DFKI GmbH,
Germany
drechsler@uni-bremen.de

ABSTRACT

Crowdsourcing platforms connect companies with heterogeneous users to create innovation ecosystems. However, platforms often have difficulty keeping users active. User engagement – the participation, interaction, and commitment among online users engaging in collaborative activities – is crucial to the continued success of these platforms. This paper presents a new approach to predicting whether a user will engage with online idea crowdsourcing platforms as a short-term or long-term user, applying a machine learning model that boasts 96% accuracy. By utilizing Explainable Artificial Intelligence (XAI)-SHapley Additive exPlanations (SHAP), we propose a framework for future research into user engagement patterns and trends across different contexts. This framework can assist platform administrators in recognizing and rewarding valuable users, ultimately leading to the lasting success of online idea crowdsourcing platforms.

KEYWORDS

User Engagement, Crowdsourcing, XAI, SHAP, Machine Learning

1 INTRODUCTION

Online collaborative communities allow people to interact, collaborate, share knowledge, solve problems, co-create value, and innovate in virtual environments. Organizations harness this through the use of crowdsourcing platforms, using various approaches and technologies to involve outside individuals in performing their tasks [1]. Applications that enable crowdsourcing such as Amazon Mechanical Turk (MTurk) incentivize individuals through monetary rewards for their contributions. However, there are also successful crowdsourcing platforms that rely solely on volunteer efforts [2]. For volunteer-based crowdsourcing platforms to be successful, users must participate in projects over the long term, thus minimizing the costs of new recruitment. User engagement is a crucial factor for the success of online crowdsourcing platforms as it influences the quality and quantity of user contributions, fosters collaboration and community building, facilitates knowledge sharing and learning, promotes long-term commitment, and drives platform growth and reputation [1]. Long-term user engagement indicates user loyalty and is essential for sustaining the competitiveness and relevance of crowdsourcing platforms [3]. However, identifying long-term user engagement and associated behavioral patterns is challenging [4] because of data availability and reliability issues, diversified and evolving user behaviors, and the subjectivity of user engagement measurements. Overcoming these challenges requires combining data analysis techniques and a deep understanding of the platform and user engagement data. This paper addresses this challenge by developing a method to determine long-term user engagement. We use a machine learning (ML) model for long-term user engagement classification and SHAP [5] to visualize the most important contributors to user engagement in an online idea crowdsourcing platform. Categorizing users into short-term and long-term users provides insights into their behavioral patterns, preferences, and engagement levels. For example, short-term users may require incentives to convert into long-term users, while long-term users may benefit from loyalty rewards or exclusive offers. Classifying users into long-term and short-term users can help platform administrators design better user-centric strategies to attract and retain valuable users.

2 USER CLASSIFICATION

The data for this analysis were collected from 22 large and medium-sized international companies that initiated crowdsourced innovation projects between 2011 and 2016 [6]. The projects generated 34,378 comments, 17,599 suggestions, 9,406 media uploads and 43,183 votes by users. Each project had two main phases: a suggestion phase and a voting phase. Users could comment and upload media files during these two phases. These phases can be repeated several times. The projects in our sample consisted of three to eight task-based phases and usually lasted for four to six months.

To define short-term and long-term users, we combined the different instances of user engagement such as suggestions, comments, media uploads, and votes for each project to create a record of each user's activities. The output label (Y) is determined based on each user's activities. The activity score for the i-th user is calculated as follows:

$$\alpha_i = \log(s_i + c_i + v_i + ep_i + em_i + epr_i) \tag{1}$$

where $\alpha, s, c, v, ep_i, em_i, epr_i$ respectively denote the activity score of the i-th user, the number of suggestions, the number of comments made, the number of votes cast, the number of phases involved, the

MuC'23, 03.-06. September 2023, Rapperswil (SG)

number of active months and the number of projects involved. Y defines the output label for an individual user in our dataset and is based on the following definition (2).

$$Y_i = \begin{cases} Short - Term & \text{for } 0 \leq \alpha_i < 0.5 \\ Long - Term & \text{for } 0.5 \leq \alpha_i \leq 1 \end{cases} \quad (2)$$

We determine user engagement by classifying users as short-term or long-term based on equation 2. Each user in our training dataset gets either the short-term or long-term label based on our use of three ML classifiers: a logistic regression classifier, a random forest classifier, and a gradient boosting classifier. Each classifier uses the same inputs, consisting of the following features:

- Number of months the user remained engaged
- Number of activities of the user on weekdays
- Number of activities of the user on weekends
- Number of projects the user remained engaged
- Number of phases the user remained engaged
- Number of peer activities in the project
- Length of the project in days
- Number of comments by the user
- Number of suggestions by the user
- Number of votes by the user

Based on these features, each observation consists of a vector shaped $(10, 1)$ with 10 features and one label Y (i.e., short-term or long-term).

3 RESULTS

We split the dataset into training (70%), testing (15%) and validation (15%). As Figure 1 shows, our predictions yield an accuracy of 96% using the random forest classifier, while the logistic regression classifier yields 92% accuracy and the gradient boosting classifier yields 87%.

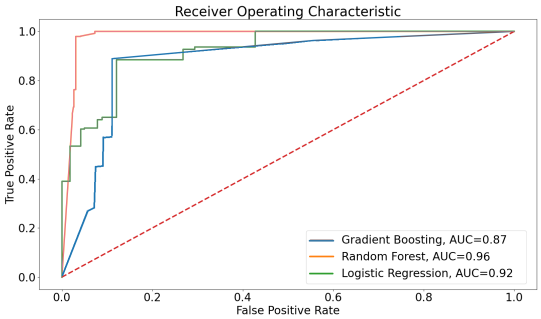

Figure 1: Receiver Operating Characteristic Curve Comparison of Different Classifiers for Short-Term vs. Long-Term User Classification

Additionally, we identified the most critical features for determining long-term user engagement through XAI-SHAP analysis of the random forest classifier. Figure 2 shows the essential set of features in predicting the long-term users in the crowdsourcing platform. Activity on weekdays and the number of engaged months are the most salient features for long-term user prediction, while the number of a user's engaged projects on the platform and the peer engagement of the projects are also significant factors in the prediction of user engagement behavior.

Because we have unlabeled (unsupervised) data and perform the data labeling ourselves, we need to deeply analyze user engagement and calculate several features that require intensive data mining. Once trained on such a dataset, the ML model can provide a framework and be easily adapted and applied to other datasets of crowdsourcing platforms. The feature selection of the ML model is not only based on the user activity score but also captures the different dynamics of a project, such as the length of the project, the number of phases, and peer engagement in the project. The XAI-SHAP feature importance reveals insights about both user- and project-based factors that influence user classification. The model will be improved in the future by optimizing the feature selection of the ML model to improve its accuracy and by predicting the user engagement level using the activity scores during different phases of the project with time series analysis. Forecasting potential short-term and long-term users can help managers motivate and incentivize valuable users to retain them.

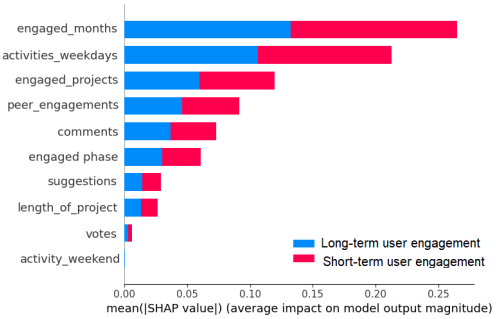

Figure 2: Feature Importance for Long-term vs. Short-term User Classification

4 CONCLUSION

This paper proposes a novel approach for categorizing users in online idea crowdsourcing platforms into short-term and long-term users. We compared the logistic regression, random forest, and gradient boosting classifiers to distinguish short-term from long-term users. The random forest model achieved the highest classification accuracy of 96%. We also used XAI-SHAP to assess the significance levels of the selected features. Our approach proposes a framework for user classification with the ML model and XAI. With this proof of concept, we will improve our work by optimizing the feature selection process and forecasting user engagement levels. Understanding the factors that contribute to long-term user engagement can help crowdsourcing platforms to design user retention strategies and enhance user-centered experiences.

REFERENCES

[1] Triparna De Vreede, Cuong Nguyen, Gert-Jan De Vreede, Imed Boughzala, Onook Oh, and Roni Reiter-Palmon. 2013. A theoretical model of user engagement in crowdsourcing. In *Collaboration and Technology: 19th International Conference, CRIWG 2013, Wellington, New Zealand, October 30–November 1, 2013, Proceedings 19*. Springer, 94–109.

[2] Andrew Mao, Ece Kamar, and Eric Horvitz. 2013. Why stop now? predicting worker engagement in online crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 1, 103–111.

[3] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, and Bo An. 2022. Prefrec: preference-based recommender systems for reinforcing long-term user engagement. *arXiv preprint arXiv:2212.02779*.

[4] Xiao Ma, Lara Khansa, and Sung S Kim. 2018. Active community participation and crowdworking turnover: a longitudinal model and empirical test of three mechanisms. *Journal of Management Information Systems*, 35, 4, 1154–1187.

[5] Colton Ladbury et al. 2022. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Translational Cancer Research*, 11, 10, 3853.

[6] Lars Hornuf and Sabrina Jeworrek. 2023. The effect of community managers on online idea crowdsourcing activities. *Journal of the Association for Information Systems*, 24, 1, 222–248.