Submissions to a COVID-19 Data Science Challenge: The Role of Skills and Platform Engagement

SABINE BRUNSWICKER, MAI ELKADY AND FENY PATEL, Purdue University[1]

1. INTRODUCTION

Firms, governments, and researchers increasingly turn to crowdsourcing platforms like Kaggle and Topcoder to tap into the collective knowledge of data scientists to solve complex societal challenges using data science (Boudreau & Lakhani, 2013). The COVID-19 pandemic of 2020 is one of such challenges. For example, in 2020 one of the COVID-19 challenges hosted on Kaggle asked data scientists from around the world to create data science models that identify emerging healthcare topics from a large amount of scientific articles about COVID‑19 (Vermicelli et al., 2021). Prior literature suggests that individuals form their intent to participate in a COVID-19 data science challenge based on different motives, including prosocial ones (Sun et al., 2020). Further, research suggests that skill-level increases the likelihood that this intent translates into a submission (Boudreau et al., 2016). However, it is unclear if an individual's engagement with the platform is a good predictor of making the first submission. Data science platforms offer various features such as forums, email notifications, database access, cloud-based code editors that allow different degrees of platform engagement. Such engagement ranges from perceptual engagement (i.e. viewing, reading tasks, posts) to action-oriented engagement (i.e. querying data from a database). But do skills and different degrees of platform engagement predict whether an intent translates into a submission? In this paper, we examine these further using survival modeling, clustering, and regularized logistic regressions.

2. CASE SETTING, DATA AND METHOD

To achieve our goal, this study uses data collected during a COVID-19 data science challenge launched on the IronHacks platform, a new Kaggle-style data science platform that offers user data science resources like Juptyerlab, a BigQuery integration (RCODI, 2021). In this COVID-19 data science challenge organized by Purdue University, participants used longitudinal social movement and COVID-19 case data to predict the foot traffic at more than 1804 places in the county of Tippecanoe (RCODI, 2020). In total, out of 450 registered platform users, 75 users had registered for the COVID-19 Data Science Challenge, and 63 had filled in a survey for self-reported skill data. The models presented in this paper use granular trace data about the users' interaction on the platform's key components (RCODI, 2021), including its forum, task and rule pages, workspace (a JuptyerLab), and a database (BigQuery). From those data we constructed measures reflecting different forms and degrees of user's platform engagement: We capture perceptual engagement (e.g. like viewing posts, viewing calendar, viewing rules, opening emails). We also measure users' information processing that goes beyond perception and measures action (such as launching the workspace to run the Jupyter Notebook). One set of features focuses on the quantity of platform engagement (eg. how many times a user does a certain action), and another one set measures the time passed until a user engages in a particular perception/action (e.g. how many days elapsed between registration and the first time he/she takes a certain action).

[1] authors in alphabetical order; all authors contributed equally

Using this data, we implemented a range of machine learning models. To understand the temporal pattern of platform engagement we ran a survival model (Johnson, 2018) with a Kaplan–Meier estimator using the Python package lifelines (Davidson-Pilon et al., 2021). To profile users with respect to their skills and platform engagements we used a combination of K-means and Ward clustering by feeding the K-means algorithm with the centroids derived from Ward's agglomerative hierarchical clustering (Oyelade et al., 2010). To build predictive models of user submission, we ran two sets of logistic regressions using the Scikit learn package (Scitkit-learn, 2021) and the Statmodel package (Statsmodels, 2021). We used l2 regularization to prevent overfitting, especially when having a small number of samples.

3. Results

3.1. Survival patterns: The temporal patterns of disengagement

In our survival model we focused on a generic temporal platform engagement measure measuring the duration between a user's registration and the last time he/she loads the app. In other words, it measures the date when they disengage and stop using the platform (the time of their "death"). We utilized a censoring date that was close to the submission's deadline. In Fig 1 we see a sharp dip in platform engagement after day 1 for users who didn't submit, suggesting that the decision to submit is made early in the process of engagement. For those who submitted we see a small dip after day 10, suggesting that submitters are engaging more consistently with the platform.

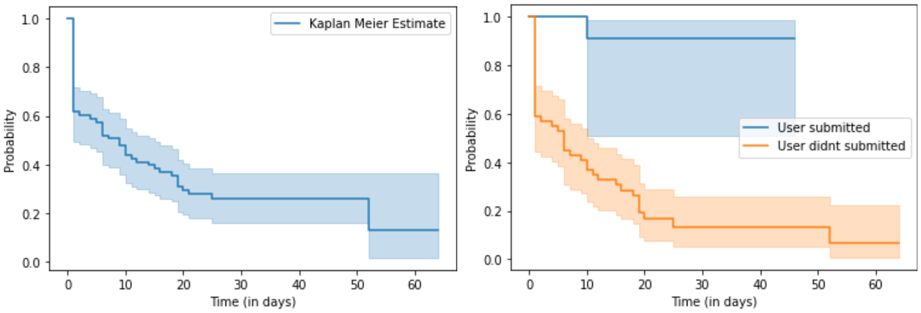

Fig 1: Survival model (Left total sample, and right: submitters versus non-submitters)

3.2. Cluster analysis: Skill and platform engagement profiles did

To better understand how skills and different degrees of platform engagement predict submission, we first clustered our participants with respect to skills and platform engagement.

Skill profiles: When we clustered the participants on their overall skill-levels, we identified a few highly skilled outliers. After removing the outliers, we clustered the remaining 55 participants using a combination of K-means and Wards. This analysis gave us two clusters with sizes of 17 and 38. Overall, cluster 1 showed participants with a higher skill set, especially in Python, R, Java, C/C++ and Bash, than cluster 2. The number of participants that submitted from cluster (higher skill set cluster) were 4 and this number was 7 for cluster 2, suggesting that skills correlate little with submissions.

Platform engagement profiles: When clustering the 75 users using 15 frequency-oriented platform engagement features, we obtained two different clusters of sizes 7 and 68. Interestingly, every participant in cluster 1 submitted. This showed a unique behavioural pattern of high platform engagement across perceptual and among most of the submitters. For further observations, we clustered the second cluster (size 68) separately to identify distinct subpatterns. We obtained three subclusters (cluster 2a,2b,2c) with sizes 15, 5 and 48 and respectively. Participants in cluster 2c showed a very low engagement profile. Cluster 2a and 2b both showed moderately high platform engagement but differed in their nature of their platform engagement. Participants in cluster 2b were more actively engaging with the data and the platform features: e.g. they frequently queried BigQuery prior to the release of the task. The number of submitters in the cluster 2a and 2b were 1 and 3 respectively.

3.2. Predicting submissions: Frequency versus temporality of platform engagement

To build predictive models of user submission, we ran two sets of logistic regressions. First, we focused on the frequency of perceptual and action-oriented platform engagement to model whether a user submits or not. In our experimental setup, we splitted the data into training and testing (with ratio 0.7-0.3 respectively), and ran the training on five different splits, and averaged the results across those five splits. We noticed that some features correlate better with user submission, and hence we decided to use a cut-off of 0.55 and include only features that have a correlation of 0.55 or above with user submission, and upon using those features with our model, we were able to achieve a maximum average accuracy of 0.922 for a regularization parameter of C= 100, with a true positivity rate (TPR) value of 0.714 and an false positivity rate (FPR) value of 0.0495 (see Fig. 2 for different values of C). The results of this model are even better than the model including all the platform metrics features (which at its best for C = 100 had an overall accuracy of 0.887, a TPR of 0.5 and an FPR of 0.059).

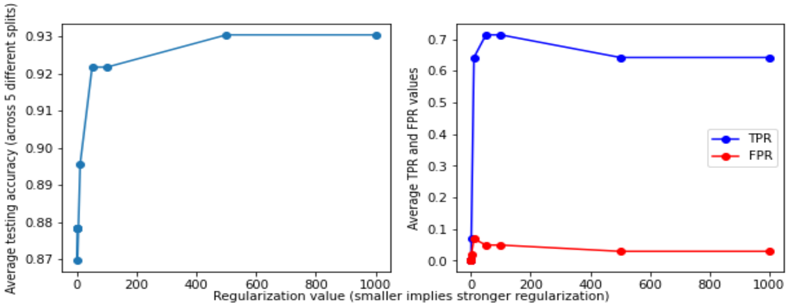

Fig 2: Average accuracy (Left) and average TPR and FPR (right) when using all features with correlation > 0.55 with user submission.

To understand the relative importance of different features, we ran an unregularized logistic regression with the features with a correlation value larger than 0.55. Only one feature had a significant independent effect (p>0.05): namely, the frequency of interacting with the BigQuery database after the task was released (odds value: 8.91). This suggests that skills and perceptual engagement (e.g. reading the task, rules, etc.) do not independently affect the likelihood of a submission. Instead, action matters: Those that interact with the data very often after the release of the task are more likely to submit.

Second, we ran a logistic regression classifier with the time stamped version features in order to predict a submission considering the timing/delay of a particular form of platform engagement after the user's registration. At a value of regularization parameter of C=100, we achieved an average accuracy of 0.945, a TPR of 0.875 and an FPR of 0.043. In addition to that we ran an unregularized logistic regression on the data to examine the independent effect of an individual temporal feature. However, all p-values were greater than 0.05. This suggests while greater frequency in engagement with the data increases the likelihood of submissions, our second model suggests that the temporality of individual perceptual and action-oriented engagement does not independently increase the likelihood of a submission. In other words, the temporality of an individual action does not explain a submission.

REFERENCES

Boudreau, K. J., & Lakhani, K. R. (2013). Using the Crowd as an Innovation Partner. *Harvard Business Review*, *91*, 61–69.

Boudreau, K. J., Lakhani, K. R., & Menietti, M. (2016). Performance responses to competition across skill levels in rank-order tournaments: Field evidence and implications for tournament design. *The RAND Journal of Economics*, *47*(1), 140–165. https://doi.org/10.1111/1756-2171.12121

Davidson-Pilon, C., Kalderstam, J., Jacobson, N., Reed, S., Kuhn, B., Zivich, P., Williamson, M., AbdealiJK, Deepyaman Datta, Fiore-Gartland, A., Parij, A., WIlson, D., Gabriel, Moneda, L., Moncada-Torres, A., Stark, K., Gadgil, H., Jona, Karthikeyan Singaravelan, … Golland, D. (2021). *CamDavidsonPilon/lifelines: 0.25.10* (v0.25.10) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.805993

Johnson, L. L. (2018). Chapter 26—An Introduction to Survival Analysis. In J. I. Gallin, F. P. Ognibene, & L. L. Johnson (Eds.), *Principles and Practice of Clinical Research (Fourth Edition)* (pp. 373–381). Academic Press. https://doi.org/10.1016/B978-0-12-849905-4.00026-5

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. *ArXiv:1002.2425 [Cs]*. http://arxiv.org/abs/1002.2425

RCODI. (2020). *COVID-19 Data Science Challenge*. www.rcodi.org/covid19

RCODI. (2021). *IronHacks Platform: Platform Documentation*. https://ironhacks.github.io/docs/

Scitkit-learn. (2021). *sklearn.linear_model.LogisticRegression—Scikit-learn 0.24.1 documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Statsmodels. (2021). *statsmodels.discrete.discrete_model.Logit—Statsmodels*. https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Logit.html

Sun, Y., Tuertscher, P., Majchrzak, A., & Malhotra, A. (2020). Pro-socially motivated interaction for knowledge integration in crowd-based open innovation. *Journal of Knowledge Management*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/JKM-04-2020-0303

Vermicelli, S., Cricelli, L., & Grimaldi, M. (2021). How can crowdsourcing help tackle the COVID-19 pandemic? An explorative overview of innovative collaborative practices. *R&D Management*, *51*(2), 183–194. https://doi.org/10.1111/radm.12443