Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing

Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing

Andrew Mao*
Harvard University
mao@seas.harvard.edu

Ece Kamar
Microsoft Research
eckamar@microsoft.com

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Abstract

We present studies of the attention and time, or *engagement*, invested by crowd workers on tasks. Consideration of worker engagement is especially important in volunteer settings such as online citizen science. Using data from Galaxy Zoo, a prominent citizen science project, we design and construct statistical models that provide predictions about the forthcoming engagement of volunteers. We characterize the accuracy of predictions with respect to different sets of features that describe user behavior and study the sensitivity of predictions to variations in the amount of data and retraining. We design our model for guiding system actions in real-time settings, and discuss the prospect for harnessing predictive models of engagement to enhance user attention and effort on volunteer tasks.

Introduction

Numerous crowdsourcing applications, such as those fielded on Amazon Mechanical Turk (MTurk), reimburse people with monetary payments for their efforts on tasks. In contrast, some successful projects rely solely on volunteer effort. These include citizen science efforts that enlist the help of large numbers of interested volunteers to provide input on solving such problems as protein folding via an online game (Khatib et al. 2011), classification of heavenly bodies from images (Lintott et al. 2008), and bird identification and tracking over large regions (McCaffrey 2005). The motivations and interests of volunteers drive the effort and attention invested in citizen science. However, little analytical work has been done to date on the engagement and disengagement of volunteer crowd workers.

We explore the challenge of learning from data to predict signals of the attention and effort that workers allocate to tasks. Such models for estimating the time and effort invested by workers are useful for understanding worker behavior and improving existing systems. For instance, predictions about engagement can help explain the influence of different interaction designs on user attention and effort at points within and across sessions of crowd work. Studies of engagement could reveal patterns of engagement for different groups of users, predict users' disengagement, direct the assignment of task sequences to volunteers so as to enhance interest, effort and attention, and measure and influence the likelihood that users will return to continue volunteer efforts at a later time. The ability to predict forthcoming disengagement of individual workers would allow systems to make targeted *interventions*, such as providing especially interesting tasks to workers at risk of becoming bored, directing support to struggling new workers, helping with the timing of special auxiliary materials or rewards, and encouraging workers to return in the long run. Data collection and modeling of engagement is also promising for the comparative study of different designs such task structures or workflows (Kulkarni, Can, and Hartmann 2012; Lin, Mausam, and Weld 2012), programs such as achievement-based badges that provide different intrinsic incentives (Anderson et al. 2013), and their influence on different types of workers.

We construct predictive models of worker engagement from large-scale usage data collected from a crowdsourcing platform. We focus on predicting that a volunteer worker will disengage within a given number of tasks or minutes, based on data about volunteers' characteristics and activities logged in histories of interaction and sensed in real time. We focus our studies on a citizen science platform called Galaxy Zoo (Lintott et al. 2008). Using supervised learning, we learn models for predicting worker engagement and evaluate them on data collected from Galaxy Zoo. The results demonstrate that learned models can successfully identify workers that are soon to disengage. We study various notions of engagement and compare the importance of different factors in accurately predicting worker engagement. Finally, given that real-world crowdsourcing systems accumulate data continuously over their lifetimes, we evaluate the amount of data and retraining needed to learn such models accurately. These studies help with understanding the factors that influence workers' engagement and provide insights about deploying predictive models in real crowdsourcing systems.

Related Work

Previous work on applying machine learning techniques to crowdsourcing has focused mainly on learning about worker quality and optimizing decisions in a crowdsourcing platform accordingly for improved task efficiency (e.g., White-

*This work was done during a Microsoft Research internship.

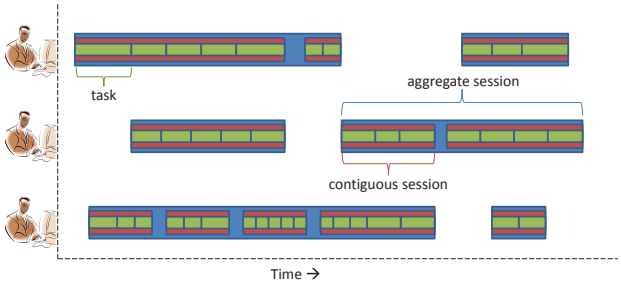


Figure 1: Model of worker sessions in crowdsourcing.

hill et al. 2009). Aggregation of the inputs from multiple workers has been used to achieve higher quality solutions for tasks (Little et al. 2010; Mao, Procaccia, and Chen 2013). Beyond simple aggregation of answers, there have emerged principled approaches to guiding crowdsourcing using decision-theoretic methods. The CrowdSynth project by Kamar, Hacker, and Horvitz (2012) introduces a decision-theoretic methodology for reducing volunteer effort while maintaining accuracy by integrating the efforts of machine vision with human perception and computing the value of acquiring additional information from workers. Efforts on TurKontrol (Dai, Mausam, and Weld 2010; 2011) provide mechanisms for choosing different workflows in a crowdsourcing system. Beyond analyses of worker quality, research on the behavior of workers in crowdsourcing platforms include observational studies on task prices, task completion time, worker availability (Ipeirotis 2010), worker incentives (Kaufmann, Schulze, and Veit 2011), and on implicit and explicit motivations of workers (Rogstadius et al. 2011). Understanding, sustaining and improving worker engagement has been mentioned as a future challenge for the crowdsourcing community (Kittur et al. 2013).

Apart from crowdsourcing, studies have shown that worker attention is related to other qualities of collective intelligence. Huberman, Romero, and Wu (2009) show that video upload activity on YouTube strongly depends on the number of views of previous videos. Kittur and Kraut (2008) find that the quality of articles on Wikipedia critically depends on the activity of numerous editors and their method of coordination. Cosley et al. (2006) describe how online communities can produce quality contributions, and gives several examples where various methods of mediating worker contributions have succeeded and failed. Beyond efforts in crowdsourcing, there have been several methodologically related studies of user attention in the context of web browsing. These efforts include work by Adar, Teevan, and Dumais (2008) that examines patterns of browsing across web browsing sessions and by Sculley et al. (2009) that explores supervised learning for making predictions about individual behavior on the web. In aggregate, the literature suggests that the effectiveness of crowdsourcing is significantly influenced by the state of worker attention.

Data, Outcomes, and Model

We consider crowdsourcing settings where workers complete a series of *tasks* over time. These settings can include tasks on paid crowdsourcing platforms or volunteer efforts

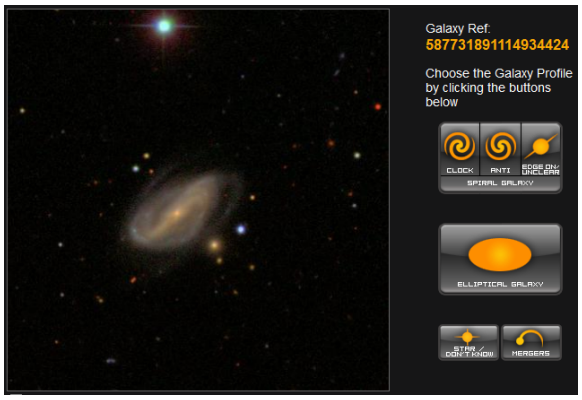


Figure 2: The Galaxy Zoo 1 classification interface.

such as commonly seen with citizen science tasks. A task is the smallest indivisible unit of work that can be completed, e.g., a single classification in a citizen science system or a human intelligence task (HIT) on Amazon Mechanical Turk (MTurk). We consider sessions of a worker on a crowdsourcing platform to be the periods of time that workers spend engaged with the platform. Workers complete multiple tasks over the course of a task-centric *session*. The progress of a worker can be interrupted for various reasons. Short-lived demands for attention such as bathroom breaks or brief conversations divide a sequence of contiguous tasks into *contiguous sessions* of uninterrupted work, divided by short breaks where workers intend to return to the task. Workers can also decide to stop working for longer periods of time or end their work for a variety of reasons; these longer pauses in activity divide the activity into *aggregate sessions*, comprised of one or more contiguous sessions.

Contiguous and aggregate sessions may have different properties in terms of the engagement of a worker. Workers are likely to maintain the cognitive context of previous tasks for contiguous sessions that start soon after the end of the prior session. Workers starting a new session after the end of an aggregate session can be assumed to return without such mental context. Because engagement within contiguous and aggregate sessions may have different implications for the crowdsourcing platform, we study them separately.

Figure 1 shows a visual representation of worker activity over time under these session definitions. Each inner segment (green) represents a task. Workers may complete tasks at different rates and the width of the segment is the length of time used to complete the task. Groups of tasks divided by brief interruptions comprise contiguous sessions (red). A sequence of one or more contiguous sessions defines an aggregate session (blue). As shown in the figure, individual workers may differ in terms of the frequency of their contiguous and aggregate sessions, the amount of time they spend in sessions, and the number of tasks they perform.

Galaxy Zoo as Testbed

Galaxy Zoo (Lintott et al. 2008) is a citizen science project that began in 2007, harnessing the power of many to classify images of galaxies from the Sloan Digital Sky Survey (SDSS) via the internet. Volunteer citizen scientists (work-

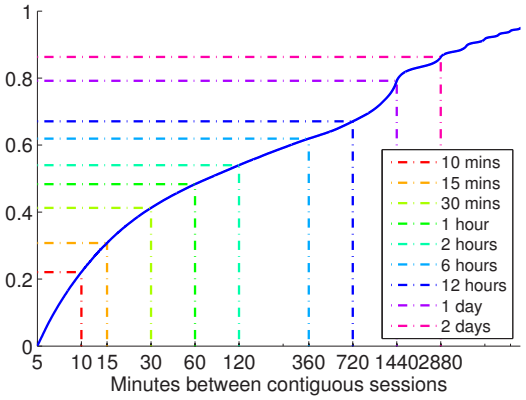


Figure 3: Cumulative distribution of inter-session times on Galaxy Zoo.

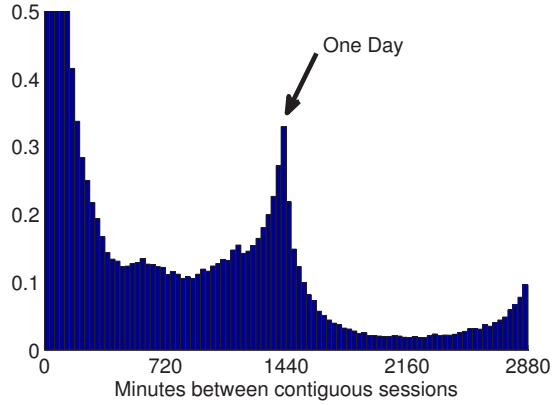


Figure 4: The distribution of inter-session times on Galaxy Zoo over a two-day period.

ers) engaging with Galaxy Zoo are asked to evaluate the morphologies of galaxies in the survey. To date, volunteers have examined nearly a million SDSS images. Currently in its fourth iteration, Galaxy Zoo is one of the longest running, most publicized, and most established examples of an unpaid, volunteer crowdsourcing system.

We study data about task completion from the first version of Galaxy Zoo. In that version, workers are shown a picture of a celestial object, and press one of six buttons to classify the object into categories such as an elliptical galaxy, spiral galaxy, or other type of object (See Figure 2). The dataset collected from the Galaxy Zoo system enables a large-scale study of engagement of workers in crowdsourcing platforms. The dataset includes 34 million votes collected from 100,000 participants about 886,000 galaxies.

Worker Behavior

The tasks and associated patterns of interaction on Galaxy Zoo are nicely captured by the representation of contiguous and aggregate tasks: each new effort at completing a classification represents the completion of a new task, which can be as short-lived as a couple of seconds. Workers complete many tasks over time, represented as one or more sessions of work divided by breaks. Some workers spend a great deal of time on the site; one worker classified nearly 12,000 galaxies in a single session, while another spent more than 17 hours making contributions. In both of these cases, no break was longer than 30 minutes.

We define the end of a contiguous session as a break of more than 5 minutes, since it is unlikely for a worker to spend this amount of time on a Galaxy Zoo task without activity. With this definition of disengagement for contiguous sessions, the average amount of time spent on each Galaxy Zoo task is 9 seconds with a standard deviation of 21 seconds.

To define a disengagement criteria for aggregate sessions, we study the distribution of the time it takes for workers to return to the platform after disengaging for more than 5 minutes (end of a contiguous session). Figure 3 shows the cumulative distribution of time between tasks when pauses are greater than 5 minutes. As displayed in the figure, many contiguous sessions are followed with a new session from the same worker in a few hours. Indeed, 41% of workers start a new contiguous session within 30 minutes of the end of their previous session. Since, 30 minutes may still preserve the context of the earlier contiguous session, we admit adjacent contiguous sessions that are less than 30 minutes apart into a single aggregate session. In the rest of the paper, we use breaks of lengths **5** and **30** minutes as the definitions of disengagement from contiguous and aggregate sessions, respectively. In Galaxy Zoo, a worker completes an average of 81 tasks ($\sigma = 146$) and spends on average 693 seconds ($\sigma = 938$) in a contiguous session. On average workers complete 135 tasks ($\sigma = 233$) within an aggregate session and the average time spent is 1629 seconds ($\sigma = 2282$). These composites of task completion and engagement times naturally resemble power law distributions.

The interval distribution shown in Figure 3 shows several aspects of the engagement behaviors of workers. Although the distribution reveals a power-law taper, visible jumps appear in the distribution at around one day, with smaller jumps at two days, three days, etc. This suggests that the longer breaks between worker sessions are not smoothly distributed, which suggests that workers have strong patterns of engagement. Indeed, for some noticeable fraction of workers, there is a high probability of returning at the same time each day—these workers have a relatively predictable schedule for investing time. This trend is more visible in Figure 4, which displays a histogram of such times for periods up to two days, using a linear scale on the time axis. Much of the mass is concentrated in the period of several hours shortly after completing a task. However, the exponential decay of return rate after completing the task is interrupted by a jump leading up to the one-day mark. If the worker does not return within one day, the distribution is similar for the second day. However, the marginal probability of returning is much lower for returns on later days. We shall now turn to analyses using machine learning and inference.

Instance Generation

We model the problem of predicting worker engagement as a binary classification problem. Each interaction of a worker

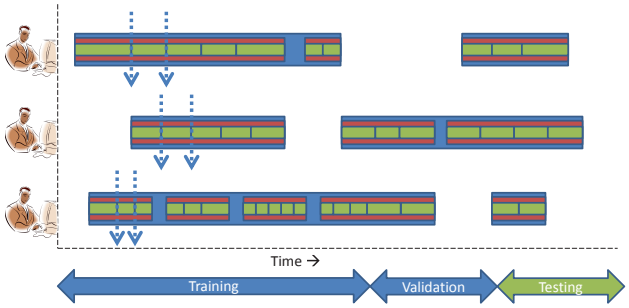


Figure 5: Instances created as cases for learning from Galaxy Zoo dataset. An instance is created for each task that each worker has completed, including features describing the historical behavior of the worker and the system.

with a task becomes an instance that is considered as a case in a library of events for training and testing predictive models. We define for each instance a set of features describing the state of the worker's interaction (including the worker's historical behavior). The label on the outcome for each instance is assigned by observing the state of the worker's engagement. We define the prediction challenges as follows:

> *Given the current state of a worker's session, will the worker stop participating within a given* **number of tasks** *or* **minutes of time***?*

If the condition defined above holds for the current state of a worker, the corresponding instance is assigned a positive label for either the time-based or task-completion versions of the prediction challenge. Otherwise, the instance is assigned a negative label.

Figure 5 shows a graphical depiction of the definition of instances. When each worker finishes a task, we create a data instance capturing the state of the worker's session, a set of features about the worker's recent and past activities (described in more detail below), and a corresponding label about engagement. The dataset consists of all worker-task interactions. We use different portions of this data to train, validate, and test learned models for predicting engagement.

The data extraction and analysis is complicated by the temporal nature of the activities and outcomes. Since each instance includes historical information about the behavior of a worker and the system, careless ordering of the data could lead to potential inconsistencies where instances in the training set contain information about instances in the test set. To avoid this potential confounding, we select training and validation instances that come strictly before test instances, as shown in Figure 5. This methodology mimics how predictions would be used in practice: training and validation instances would be constructed from existing data, and predictions would be used to make decisions about subsequent worker sessions at run time.

Labels

In the context of the instances above, we can employ several labels on outcomes that describe disengagement over time. We may be interested in how soon a worker will stop working, or how many more tasks they will perform. Some outcomes may be easy to predict. Other outcomes may be less easy to predict but make for more valuable predictions.

The challenge of predicting whether a worker will stop working in the next 30 seconds is significantly different from the challenge of predicting whether the worker will stop within 30 minutes. These predictions would likely be used in different ways in the operation of a crowdsourcing system. The former outcome has very few positive instances, and training a classifier for such biased data sets can be challenging. We focus on binary outcomes on disengagement—on the challenge of predicting whether the worker's session will end within a given amount of time or number of tasks, and report our findings in the following section.

Features

As shown in Figure 4, workers may have strong patterns of engagement, including recurrent activities with time-of-day resonances. Such patterns of effort can inform the construction of features. We formulate features and group them under three general categories.

Task-Based Features. Without considering behavioral changes over time, workers may be affected simply by the tasks that they observe; this assumption underpins many worker/task latent-feature models in crowdsourcing (see Raykar et al. 2010 for an example). The Galaxy Zoo team shared with us anecdotal evidence suggesting that workers tend to have a longer session if the first few galaxies they see are interesting, high-quality pictures, rather than the more common less interesting or low-quality galaxy images. We can evaluate this objectively by computing features that capture worker behaviors in response to sequences of difficult or banal tasks, based on the activity of other workers. These features include those based on use of an estimate of the running difficulty of the last X tasks, computed by considering differences in votes on objects by others. We summarize differences in votes on objects via computing the entropy of a set of answers.

Session Features. We also consider attributes that characterize workers' activities within the current session. We consider statistics around the number of tasks completed in the current session versus completed in typical sessions for each worker. We also compute statistics about the *dwell time*, capturing the amount of time spent on each task, and the worker's *vote entropy*, which represents the diversity of workers' classifications. We believed these statistics could serve as signals of a worker's attention to tasks at hand. For example, a running average of dwell time as compared to the average dwell for sessions can measure whether the worker is starting to pay less attention or struggling on a given task. Similarly, a worker providing a set of votes with low vote entropy on a spectrum of randomly sorted tasks may be selecting the same classification for many tasks in the absence of deep analysis, and thus paying less attention than someone who is providing input that is better matched to the distribution of cases. All of the features mentioned can be derived from behavior in the current session regardless of the worker's histories or habits. We compute these features for

both contiguous and aggregate sessions as the characteristics may be different.

Worker Features. We can also compute multiple features that characterize workers based on their history and habits. Such features are a rich source of information for learning to predict future engagement of individual workers. These features include the following classes:

- **Summary features.** These features include the typical time spent on tasks, number of past sessions, and average time spent on sessions. These features implicitly distinguish among segments of the worker population.
- **Start-/end-time features.** Features build on periods of time when workers engage with the system, including comparison of the current time of day to the typical time of day that the worker has started or ended a session in the past.
- **Session history features.** Features describing the worker's behavior in aggregate sessions, including the number of short breaks that are taken and the length of contiguous sessions.
- **Inter-session features.** These features capture information about the period of time (gap) since the worker's last session and how this compares with past gaps.
- **Dwell time features.** Features on the amount of time that the worker spends on tasks, including consideration of statistics of different running averages compared to previous computed averages over the worker's history.
- **Session task features.** These features include a set of compound features that compare the worker's task-related statistics on the current session with the mean statistics observed on past sessions, including number of tasks completed and amount of time spent.

We compute statistics on features for the complete history of a worker and also for the most recent history (i.e., last 10 sessions) to identify behavioral changes. The worker features also implicitly include session features, as they compare a worker's current session to longer histories of behavior. Overall, we computed nearly 150 features for our data instances.

The features that we compute do not explicitly encode domain-specific knowledge about the specific task of classifying galaxies. For example, no feature depends on the results of any automated classification of galaxies, or a prior distribution of the types of galaxies. While using domain-specific features may improve predictive performance, we focused on how well we can detect worker engagement using features that are applicable to numerous other types of tasks. We believe that the methods can be generalized to similar crowd work settings.

Evaluation

We seek the construction of statistical models that can predict that a worker will disengage within some horizon of time or tasks. We generate our datasets for experiments on these predictions from the three months of Galaxy Zoo data using the methodology described in the earlier section. We remove from consideration workers for whom we observed little activity (less than 10 contiguous sessions). The generated data set consists of over 24 million instances, corresponding to each task that was performed by the set of workers that we considered. For each experiment, we randomly sample 500,000 training instances, 250,000 validation instances, and 250,000 test instances, preserving temporal consistency per above. This approach ensures that all of the methods use the same amount of data when possible. Unless otherwise noted, we used the complete set of features in the experiments.

Predicting the instances described below typically results in biased data sets, containing very few positive instances where users disengage. As a result, we consider the measure of area under the receiver-operator characteristic curve (AUC) to evaluate the relative performance of different classification algorithms. The AUC measure can be interpreted as the likelihood that a classifier will distinguish a randomly selected positive instance from a randomly selected negative instance. A random classifier that assigns each instance the prior probability of the dataset has an AUC of 0.5, and a classifier that can always distinguish positive from negative instances has an AUC of 1.0. The AUC is invariant to the prior distribution of labels in different datasets, which can be highly skewed. We additionally measure the log-loss reduction (LLR) achieved by each classifier as compared to the random classifier as a measure of the accuracy of probabilities assigned to each instance by the model. Higher positive log-loss reduction values predict more accurate probability estimates. In all of our experiments, classifiers with higher AUC values showed higher log-loss reduction metrics. For simplicity of presentation, we report AUC values for the experiments, since they provide a robust metric for comparing the success of predicting different outcomes.

For each classification task, we performed a validation phase to learn the best classifier. We explore the predictive power of models constructed with boosted decision trees, linear SVM, and logistic regression for predicting the outcomes described below. For each procedure, we normalized the data and performed parameter sweeps using the validation set. We created a single best classifier for each task by identifying the procedure and parameters that performed the best on the validation data. We report the results of the final models on the test set below. In our experiments, boosted decision trees consistently outperformed SVM and logistic regression on the validation set, and thus was used to train all of the final classification models.

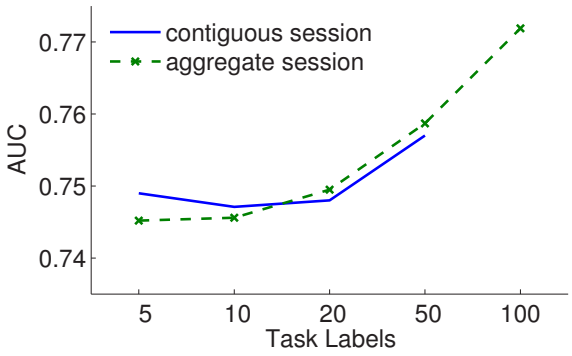
Outcomes of Interest

For each instance in our dataset, we are defining outcomes according to the definition below:

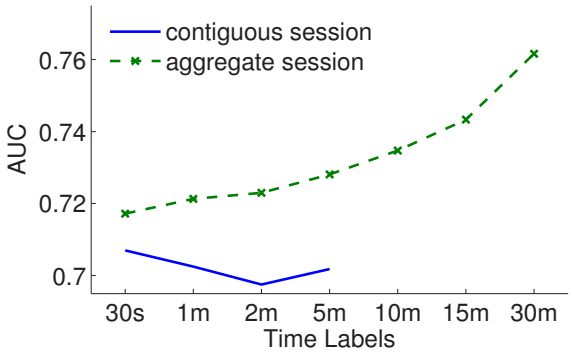
> Does the worker's current (**contiguous / aggregate**) session end within (X **tasks /** Y **minutes**)?

For example, if the particular outcome of interest is whether the aggregate session ends in 20 tasks, then a positive instance indicates that the worker will stop within the next 20 tasks and that they will not return for at least 30 minutes.

This definition is quite general; it includes definitions of disengagement outcomes based on different session defini-



(a) Predicting outcomes defined in terms of number of tasks.



(b) Predicting outcomes defined in terms of time.

Figure 6: Prediction performance with different outcomes, using all features.

tions. The closeness to disengagement can be defined based on the number of tasks or the amount of time, and the degree of closeness can vary with different X or Y values. While some outcomes may be more easily predictable than others, specific predictions may be particularly useful for a domain in guiding decisions, such as interventions aimed at increasing effort allocated by volunteers. Designs for valuable target predictions and best uses of inferences will typically be domain-centric exercises. Given the range of potential uses of predictions about disengagement, we do experiments over a spectrum of outcomes.

Figure 6 shows the performance of predictions for different target outcomes, as indicated on the x-axis of the graphs. Generally, we can better predict the end of an aggregate session (where the worker does not return for at least 30 minutes) than the end of a contiguous session (the worker does not return for at least 5 minutes), especially in terms of time. As might be expected, we can better predict whether a session will end within a larger number of tasks or longer period of time than within a small period. Figure 7 shows the ROC curves for predicting the end of an aggregate session by number of tasks. The AUC monotonically increases as we increase the number of tasks.

Despite these trends, the analyses show that extreme differences do not exist among the predictability of different outcomes. We shall focus on the example of predicting the outcome that a worker will quit an aggregate session within 20 tasks since the target outcome is far enough ahead to allow for the execution of targeted interventions.

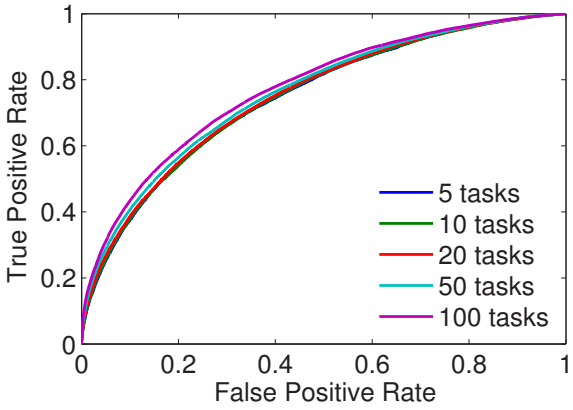


Figure 7: ROC curves of different models predicting the end of an aggregate session by number of tasks.

To put the performance of the predictive model in a usage context, consider the following: suppose that we seek to guide interventions based on identifying workers who will leave an aggregate session within 20 tasks. Using random targeting, only 14% of interventions would reach our target group. However, using the predictions on outcome to target the top 0.1% of workers likely to leave, 85% would reach our target group, a significant increase from the 14% made by random choice. This number would be 79%, 72%, 54%, and 44% by targeting the top 0.5%, 1%, 5% and 10% respectively, giving a tradeoff between accuracy and number of targets reached.

Feature Selection

The next set of experiments study which sets of features are most predictive of disengagement within a horizon. We wish to understand the accuracy of predictions as models are provided with larger sets of features. We are also interested in the relative influence of different feature sets on predicting disengagement for workers when we have small versus larger histories of interaction. For example, worker features may be more discriminatory when we have a great deal of historical information. We study the influence of quantity of historical data on prediction performance by sampling two additional test sets, consisting of data instances in the top and bottom quartiles of worker history activity by number of past aggregate sessions.

Figure 8 shows the prediction performance for small amounts of history history, large amounts of history, and for all workers for combinations of the following sets of features: task (T), contiguous session (C), aggregate session (A), and worker features (U). The results show that all feature sets individually help to predict worker engagement. However, adding worker features with session features results in a large boost in prediction performance, and the benefit of including task features is diminished. We also see that workers with larger histories are more predictable even when the models do not include worker features. On the other hand, adding features describing past behavior produces less improvement (relative to the population at large) for workers

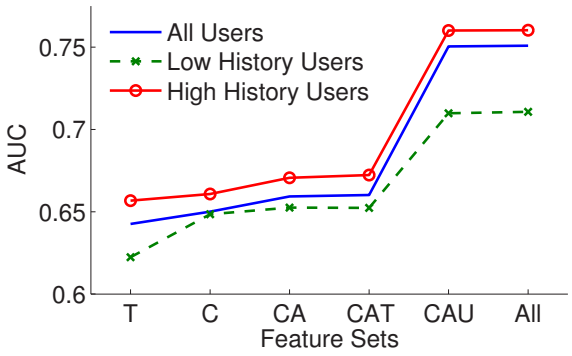


Figure 8: Model trained with different subsets of features and tested on different subpopulations.

with small amounts of history, as one would expect.

For the model trained with all of the features, the most predictive features, as measured by information gain in the boosted decision tree ensemble, are primarily about contiguous and aggregate sessions and the worker's history. The most informative feature is the average number of tasks in recent (the last 10) aggregate sessions, followed by the number of tasks over the worker's entire history, and over the last 10 contiguous sessions. Other informative features compare past behavior with recent behavior (e.g., difference of the average number of tasks done in an aggregate session in the entire history versus completed more recently) and features about the current session (e.g., average dwell time in the current session).

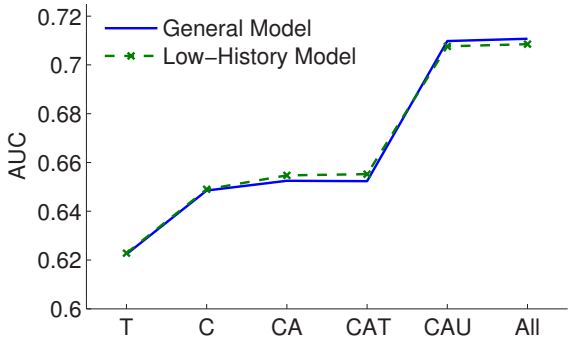
Worker Groups

Our results in the previous section suggest that the performance of predictive models depends on the specific worker subgroup at focus of attention. Hence, we consider whether we can gain a further advantage by training a prediction model for only specific subsets of workers. For example, we may be particularly interested in using targeted interventions to enhance the engagement of workers with small amounts of history so as to guide them early on to becoming more involved with a specific citizen science community.

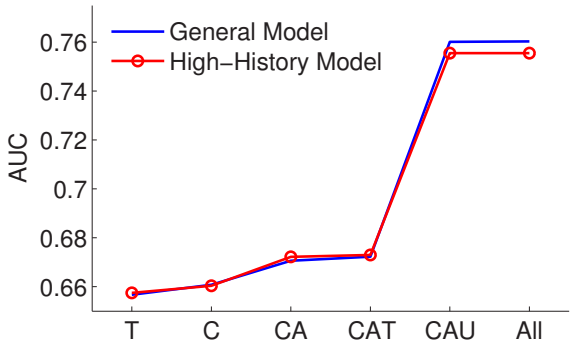
Figure 9 shows the results for predicting the engagement of workers with small and large histories when models are trained with the data collected only from each class of workers. The results show that training for specific subsets of the workers does not improve the performance of predictions. These results suggest that, when there is a large amount of data available from a crowdsourcing system to generate instances and create features, a relatively complex classifier trained on the entire dataset may generalize well to specific subcategories of workers.

Cold-Start Performance

In a typical scenario, a crowdsourcing system begins soliciting contributions on a new type of task. At the outset of the use of the system, there is little data about workers and it may be difficult to make predictions about engagement when few workers have extensive histories. A best current model may not generalize well to future predictions as



(a) Low-history workers.



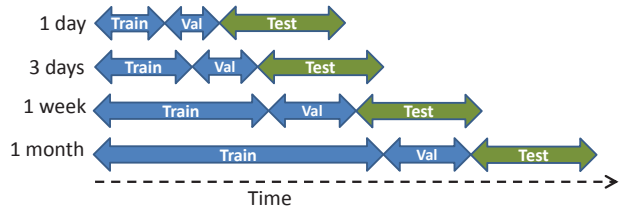
(b) High-history workers.

Figure 9: Comparison of the general model applied to a subpopulation with one trained explicitly on the subpopulation.

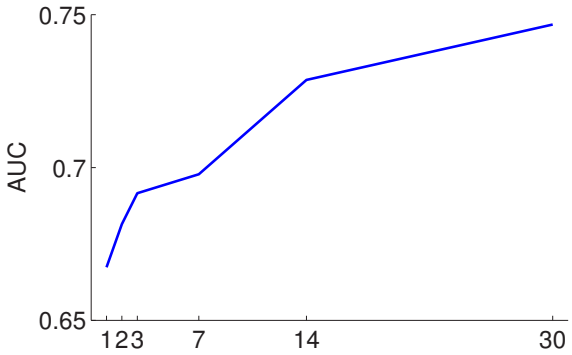
worker habits evolve and change. In our next set of experiments, we study the effect of the amount of data collected about workers on prediction performance.

Figure 10a demonstrates the approach for studying this "cold-start" problem. In each experiment, starting from the specific date when our dataset begins, we sample training and validation instances from the first day, first two days, first three days, first week, first two weeks, and first month of the system's records. Except for the first day, the amount of training data sampled stays constant between experiments; however, the data set becomes more heterogeneous and represents more diversity among workers with the expansion of the training period. The test set is always sampled from the one-week period immediately after the training and validation sets. This formulation of sampling mimics the challenge of making predictions about engagement in real time as the test set is exactly the subsequent set of worker sessions appearing in the system.

Figure 10b displays the results of this set of experiments. The figure shows that the system needs to collect almost a month of data to reach AUC of 0.75—the accuracy of the model trained with the complete data. The performance of the models improves rapidly as data arrives during the first few days of the system's life. This suggests that in early stages of deployment, it is generally important to continue training a prediction algorithm when more diverse information is collected about workers over time.



(a) Generating cold-start data.



(b) Performance over more days of training data.

Figure 10: Testing prediction performance in a cold-start setting.

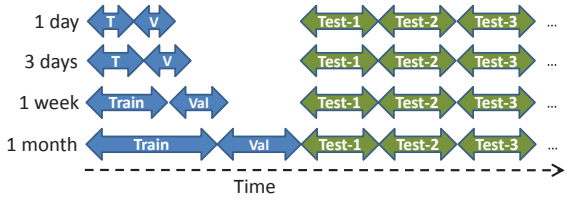
Model Generalization

Finally, we study how well a trained model generalizes for instances encountered at later times and whether the performance may diminish over time. Figure 11a shows the process for testing this problem. After training models using data sampled from the first day, first week, first two weeks and first month of the system's life, we evaluate the models on test sets for each two-week period following the first month of the system's use. Figure 11b shows the performance of the models when tested on later time segments. The results show that all models generalize well to future instances and that the performance of the models does not diminish over time. They also confirm our earlier result that models trained with data sets (of equal size) containing observations about a more diverse population consistently perform better.

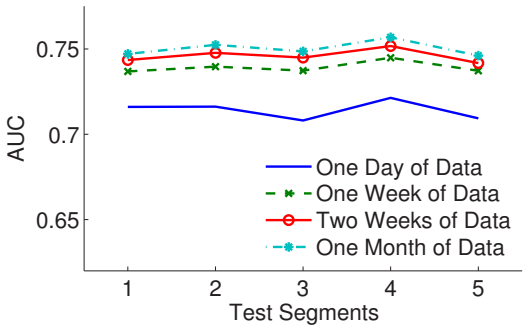
Discussion and Future Work

We presented the construction of predictive models of engagement in volunteer crowdsourcing, using data logged on the activity of citizen scientists using Galaxy Zoo. We performed several different experiments to probe characteristics of the prediction challenge. Our results demonstrate the performance of predictive models of engagement on a large-scale citizen-science platform. The trained models reached desirable performance with predicting forthcoming disengagement in different experimental conditions. Finally, we provide insights about the quantity of data needed to train models to perform well and how well the models generalize to making predictions about future instances.

We see numerous practical applications of predictive models of engagement. We expect that inferences about workers nearing disengagement can be employed in designs



(a) Generating data for evaluating generalization.



(b) Performance over future testing sets with increasing distance.

Figure 11: Testing generalization performance.

that use well-timed interventions to extend the engagement of workers. For example, it may be useful to target new workers who are about to leave a system by presenting a tutorial or a link to a discussion forum. Similarly, interventions may target workers who are struggling or losing interest by presenting more interesting tasks or by encouraging them with merit programs such as a badge programs (Anderson et al. 2013). If even only a small fraction of these workers respond to the interventions by staying and continuing to work, or returning to the platform with higher likelihood, then the platform can gain a significant benefit from predictions about disengagement.

We foresee opportunities for developing a variety of predictive models about engagement. For example, we may wish to predict if and when a worker will return after one or more sessions, based on multiple features, including traces of the worker's history of experiences with the platform. Models of engagement can be expanded to make predictions about more general notions of worker engagement, attention and effort, and they can be applied to tasks that go beyond of labeling. Beyond use on volunteer-centric tasks, we envision applications of models of engagement in paid systems. Such models may include distinctions and inferences about the joy or excitement associated with tasks, the link between intrinsic reward, payments, and effort, and leveraging of more detailed worker profiles, including demographic information and long-term histories of engagement. We hope that this work will stimulate further research on user attention, effort, and engagement in crowd work.

Acknowledgments

We thank Chris Lintott for sharing the Galaxy Zoo data, Paul Koch for assistance processing the data, and Rich Caruana for valuable discussions and feedback.

References

Adar, E.; Teevan, J.; and Dumais, S. T. 2008. Large scale analysis of web revisitation patterns. In *Proceedings of the 26th ACM Conference on Human Factors in Computing Systems (CHI)*.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2013. Steering user behavior with badges. In *Proceedings of the 22nd International World Wide Web Conference (WWW)*.

Cosley, D.; Frankowski, D.; Terveen, L.; and Riedl, J. 2006. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proceedings of the 2006 ACM Conference on Human Factors in Computing Systems (CHI)*, CHI '06, 1037–1046. New York, NY, USA: ACM.

Dai, P.; Mausam; and Weld, D. S. 2010. Decision-theoretic control of crowd-sourced workflows. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI)*.

Dai, P.; Mausam; and Weld, D. S. 2011. Artificial intelligence for artificial artificial intelligence. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*.

Huberman, B. A.; Romero, D. M.; and Wu, F. 2009. Crowdsourcing, attention and productivity. *J. Inf. Sci.* 35(6):758–765.

Ipeirotis, P. G. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17(2):16–21.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*, 1–11.

Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K.; Makedon, I.; Popović, Z.; Baker, D.; and Players, F. 2011. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* 108(47):18949–18953.

Kittur, A., and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW)*, CSCW '08, 37–46. New York, NY, USA: ACM.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318. ACM.

Kulkarni, A.; Can, M.; and Hartmann, B. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work (CSCW)*.

Lin, C. H.; Mausam; and Weld, D. S. 2012. Dynamically switching between synergistic workflows for crowdsourcing. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*.

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; Murray, P.; and Vandenberg, J. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389:1179–1189.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the 2nd Human Computation Workshop (HCOMP)*.

Mao, A.; Procaccia, A. D.; and Chen, Y. 2013. Better human computation through principled voting. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*.

McCaffrey, R. E. 2005. Using citizen science in urban bird studies. *Urban Habitats* 3(1):70–86.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11(Apr):1297–1322.

Rogstadius, J.; Kostakos, V.; Kittur, A.; Smus, B.; Laredo, J.; and Vukovic, M. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media: Barcelona, Spain*.

Sculley, D.; Malkin, R.; Basu, S.; and Bayardo, R. J. 2009. Predicting bounce rates in sponsored search advertisements. In *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining (KDD)*.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems* 22(2035-2043):7–13.