### **Checkpoint 1 Report - CPSC 6300 Hubig**

### **Introduction**

We are looking to use a machine learning model to help a crowd worker determine whether or not they should take a task based on if they would complete or not complete an individual task based on previous information and data training.

### **Summary of Data**

The datasets that we are examining are from Amazon Mechanical Turk (MTurk) and Toloka Yandex. They observe individuals and the actions they perform while working tasks, storing them as individual events, on their respective platforms. Combined they consist of 302 unique individuals contributing 12,401,997 individual actions/events towards this dataset. There is over 4 days worth of information in the MTurk dataset; whereas, Toloka has over 43 days worth of information. Both datasets have 10 columns worth of quantitative and numerical data which includes event id’s, page urls, page actions, API calls, task status, epoch timestamp, working status, person id, and two other columns that contain unusable data. Our next step was to condense our dataset and to make new columns that contained more relevant data to our problem, especially since the API calls had a bunch of useful information in the API json request.

### **Data Processing & Cleaning**

The original MTurk dataset was generated from a Turkenator Chrome Extension (plugin) that was modified to collect information necessary to report on working and invisible labor events. We noticed that the events included (user\_id, task\_id) fields that we could leverage to group events together for our final feature set; designated values (ADDED\_TASK, FINISHED\_TASK) from a subtype event field provided ranges for start/stop for a given task. We had to filter out tasks that did not include both an (ADDED\_TASK, FINISHED\_TASK) entries as this would show incomplete information. An event timecode field was provided in milliseconds which needed to be converted to a datetime format to easily identify when an event occurred; this field was important for sorting events in a chronological order of occurrence and was necessary when aggregating the duration and count totals correctly. Flattening of an event details JSON field was needed to retrieve additional data fields (task\_id, requester\_name\_company, assignment\_duration\_seconds, project.assignable\_hit\_counts, project.monetary\_reward\_in\_dollars, etc.). We calculated new fields for our feature set to include total, working, and invisible labor count and duration in seconds; adjusted monetary reward in dollars from the working + invisible labor; completed\_task (yes/no). Processing and cleaning this data was not easy because it relied upon internal knowledge about how this data was collected through the Chrome plugin, and also figuring out how to best group individual event recordings for a task was a challenge because we had to figure out when they started and stopped.

### **Data Visualization & Key Predictors**

The projected model that will use this dataset will be constructed to predict the completion status of a task, given selected features associated with a task. As described before, completion status is a binary field, where 1 corresponds to a completed task, and 0 corresponds to a task that was never finished. As such, a classification model will be used to determine this discrete outcome with two options. Figure 1 gives a glance at the distribution of completed tasks for a subset of our cleaned, task-grouped dataset.

Due to lack of familiarity with this type of dataset, an EDA had to be performed to understand how the given features correspond with our outcome, whether a task was completed or not. Some of the features were seemingly too irregular or unrelated to task completion to use in our model, like the total time allotted by a requester for the task. However, there were three very promising features that were selected as key predictors because they showed an interesting relationship when plotted against each other for classification: monetary reward for the completion of the task, number of human intelligence subtasks ("HITs") for the task, and total number of events recorded which includes both invisible labor events and visible labor events. Figure 2 has several plots showing how these key predictors relate to one another, with orange and blue legends corresponding to whether the task was completed or not, respectively.

### 

***Figure 1 Figure 2***