Homework # 1

Guanghua Wang

April 23, 2019

In this homework, I collect users who star, forks the repository: *996icu/996.ICU*. I also collect users who submit pull requests to this repository. *996icu/996.ICU* is a repository initiated by IT practitioners to "uphold the labor law and request employers to respect the legitimate rights and interests of their employees". The name 996.ICU refers to "Work by '996', sick in ICU". "996" is a work schedule that requires a worker to work from 9 a.m.–9 p.m., 6 days per week. [1].

I collect 39990 users who star *996icu/996.ICU* from 2019-03-26 to 2019-03-28. They are stargazers from the first 1333 pages. Github's API does not allow to scrape stargazers after page 1333. At 2019-04-23 22:30:00, 233,485 users star this repository. I also collect 19368 users who fork the repository as of 2019-04-22 20:56:09 and 976 users who submit pull requests as of 2019-04-22 01:25:46.

In this homework, I mainly investigate which types of workers are more likely to submit pull requests to this repository. The final data analyzed is based on 39990 users who star the repository. I drop 34 users who closed their accounts after star this repository. There 39956 workers in the final sample. Among them, 221 workers who submit pull requests.

Table 1 reports the features I use in the machine learning part. The starting date for time variables is the date when *996icu/996.ICU* was created, which is 2019-03-26 02:31:14. Comparing to users who only star the repository, users who submit pull requests have more followers, follow more others, have more repositories, open the github accounts earlier and update their repositories more recently. To sum up, workers who submit pull

[1] Check the repository for more details about this movement

requests are more attached to the github than those who only star *996icu/996.ICU*.

Table 1: Means of Workers Who Star the Repository

Variables	Submitting pull request	Not submitting pull request
Number of followers	62.769231	10.683554
Number of following	27.411765	14.695910
Number of repositories	34.733032	19.512244
Date of created	-1336.330317	-1181.541034
Date of last updated	5.438914	-11.959482
Number of Users	221	39735

I use the date when *996icu/996.ICU* was created as the starting day, which is 2019-03-26 02:31:14.

Considering the small number of users who submit pull requests, I use Logistic Regression classifier to predict users who would submit pull requests. Considering that users who may hear this repository from news reports and create github account thereafter, I create a variable to identify whether workers open github account before the creation of this repository. I also create the quadratic form of the number of followers, following and repositories and dummies for zero follower, following and repository respectively. These variables account for the nonlinear effects of number of followers, following and repositories on the probability of submitting pull requests.

To validate the prediction from my Logistic Regression classifier, I split data into two parts: one for training with 80% of observations and the other for validation with 20% of observations. The accurate score is 0.9942. However, the high accurate score is caused by large number of predicted users who do not submit pull requests. The confusion matrix in Table 2 suggests a different story from the accurate score. The machine learning algorithm I use do not predict workers who would submit pull requests well.

My algorithm need more adjustments in the future, such as utilizing locations of users.

Table 2: Confusion Matrix

	Predict 0	Predict 1
True 0	31780	10
True 1	174	1