Reproduce “Private traits and attributes are predictable from digital records of human behavior”

JING CHEN, University of Pittsburgh, Pittsburgh, PA, jic118@pitt.edu  
Yuxuan liu, University of Pittsburgh, Pittsburgh, PA, yul196@pitt.edu

KEYWORDS

data mining, private traits, Facebook likes, SVD, machine learning

Abstract

The work we will do is reproducing research results in the paper “Private traits and attributes are predictable from digital records of human behavior” [1].

1 INTRODUCTION

The motivations that we try to reproduce the result of this paper can be conclude as follow. First, following the paper instruction and reproducing the whole research process allows us to verify its result, whether the result is reproducible by the procedures given in the paper, how the difference of dataset can affect the result and whether the methods they implemented in this paper is reasonable. Second, by reproducing the study, we can put data mining theory and technical skills learned in class into practice. Also, we will try to find whether we can make some adjustments that can improve the result and provide a detailed discussion on the adjustments. We believe this work can give us a deeper understanding of knowledge in this field. Last, this problem focus by the paper is meaningful and profound, nowadays, we are exposed to social media environment, there are a huge number of easily accessible digital records of behavior on such social media, which can be used to automatically and accurately predict a range of highly sensitive personal attributes. Data mining technique skills like SVD, linear regression model, logistic regression model, cross-validation, ROC, AUC, and Pearson product-moment correlation had been used in our work.

2 DATASETS

The dataset used in our work download from the myPersonality Project[[1]](#footnote-2) database. This dataset includes three csv files, users (110K), likes (1.58 million), and userslikes (10 million). Users has the information about 110K users, including userid, gender, age, political parties, openness, conscientiousness, extraversion, agreeableness, and emotional stability. Likes has two columns, likeid and name. Users-likes has two columns, userid and likeid.

3 FINISHED WORK

The first step is to preprocess users-likes datasets, according to the support information of this paper, that requires to remove Likes which associate with fewer than 20 users, as well as users with fewer than two Likes. To be specific, we grouped the users-likes matrix by likeid and counted the userid, then we removed likeid related to fewer than 20 users. Also, we grouped the users-likes matrix by userid and counted the likeid, then we removed userid related to fewer than 2 likes. Finally, we removed users and likes which were removed above in users-likes matrix. After doing that, we got 105,262 users and 54,457 Likes. But the values of them cannot be used to create sparse matrix because they are character string. So we added indices to the reduced users and reduced likes matrixes and then joined them with users-likes matrix (left join on userid and likeid). After the join, we got a matrix with two columns userid and likeid whose values are integer. Then we created a sparse matrix (105262\*54457) which generated by a package called Matrix in R. This sparse matrix only has a small size of 87Mb (the dense matrix is about 40Gb).

Next, dimensionality reduction technique SVD has been implemented to the matrix and got two matrixes (one has 100 top SVD components, the other one has 30 components).

For the next phase, we applied cross validation on the dataset, in which, logistic regression model has been used for the dichotomous variables, linear regression model for the numerical variables.

In order to evaluate prediction performance of each model, we express the prediction accuracy of dichotomous variables in terms of the area under the receiver-operating characteristic curve (AUC). The prediction accuracy of numeric expressed by the Pearson product-moment correlation coefficient between the actual and predicted

values.

4 RESULTS AND DISCUSSION

The preliminary prediction results of numeric variables are showed in [Fig. 1](#fig2). As we can see, the highest correlation was age(r=0.42). The following are the traits “Openness” (r=0.34). However, the performance on each trait was inferior to the results given in the original paper. This chart shares the same tendency with original result chart, that means we are close to the result the paper given.

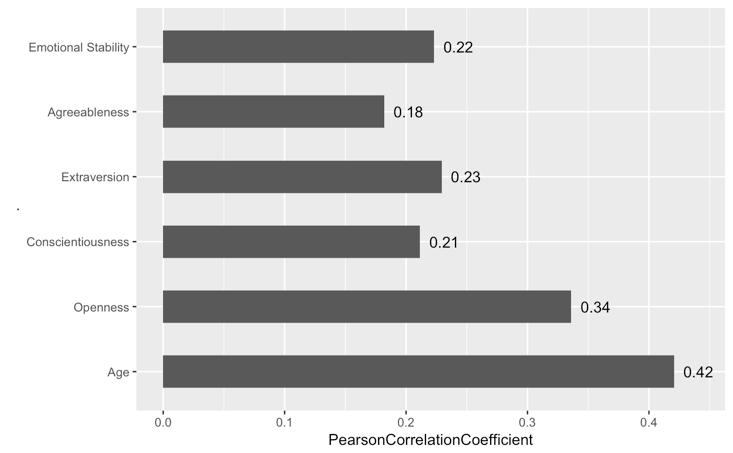


Fig. 1. Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values.

The prediction results of dichotomous variables are presented in [Fig. 2](#fig3). Gender were classified in 86% cases, Democrat and Republican were classified in 78% cases.

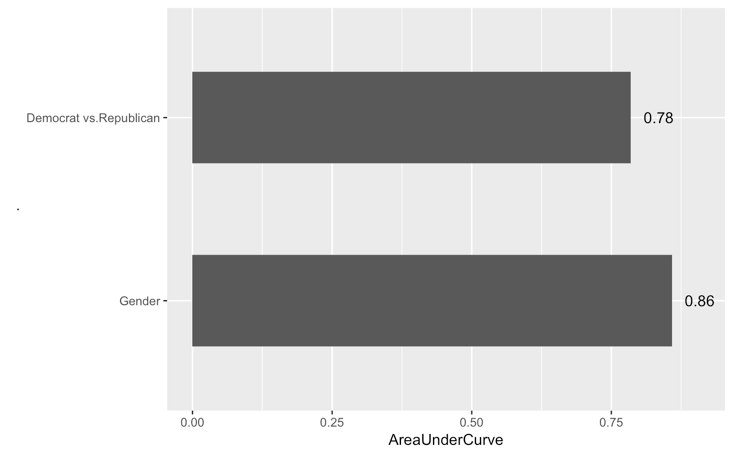


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

REFERENCES

[1] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Journal Name:ACM TRANS. GRAPH.

Volume:9

Issue:4

Year:2010

Month:March

Article Seq:39

DOI:10.1145/1234

Copyright Year:2017

Copyright Statement:rightsretained

ISSN:0730-0301

Price:$15.00

1. Data deposition: The data reported in this paper have been deposited in the myPersonality Project database (www.mypersonality.org/wiki). [↑](#footnote-ref-2)