Predicting Personal Traits from Facebook Likes --Reproducing of Michal Kosinski's Research

Chen Ruo
University of Pittsburgh
Information Science
chr87@pitt.edu

ABSTRACT

This project aims to reproduce and verify the data analysis result in Michal Kosinski, David Stillwell, and Thore Graepel's paper of "Private traits and attributes are predictable from digital records of human behavior"[1]. The original paper revealed that the personal information can be predicted with Facebook Likes. The conclusion was supported by detailed data analysis results. The reproducing work follows all the methods that were used in the original paper. Some necessary adjustments are made due to limited computing power and data accessibility. The results of reproducing work are presented and compared with original work.

CCS CONCEPTS

• **Applied computing~Document analysis**

KEYWORDS

Reproduce; prediction; digital footprint

ACM Reference format:

G. Gubbiotti, P. Malagò, S. Fin, S. Tacchi, L. Giovannini, D. Bisero, M. Madami, and G. Carlotti. 1997. SIG Proceedings Paper in word Format. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK'97)*, 4 pages. https://doi.org/10.1145/123 4

1 INTRODUCTION

In the paper of "Private traits and attributes are predictable from digital records of human behavior"[1], utilizing the 58,466 Facebook user's Like records, and the user's personal information that comes from myPersonality[2], the authors trained Logistic Regression models for all the dichotomous features and Linear Regression models for all the numeric models. For all the models, SVD (Singular Value Decomposition) is performed and 10-fold cross-validation is performed[3]. As the result of the analysis, for features including race, gender, sexual-orientation, age, etc., the model can make a prediction with pretty high accuracy.

The result of the original research is influential and profound. The power of prediction personal information from simple digital footprint can be both exiting and frightening. The reproducing and verification of the paper becomes necessary. Especially in today, with the widespread news of Facebook information sharing issue, more and more internet user would like to know that to what degree the personal information can be extracted and predicted from the digital records.

Furthermore, the reproducibility is a very important standard in evaluating the quality of the research and event the academic integrity.

In this paper, all the methods would follow the research of "Private traits and attributes are predictable from digital records of human behavior". Additionally, all the parameters would also be set the same as they were in the original research.

2 RELATED WORKS

Since this research is trying to reproduce Michal Kosinski, David Stillwell, and Thore Graepel's research that was reported in "Private traits and attributes are predictable from digital records of human behavior"[1], the original paper is the most important work that is referred. Golub GH, Kahan W's work of "Calculating the singular values and pseudo-inverse of a matrix"[4] was referred to understand and implement the Singular Values Decomposition (SVD). Koren's works of "Matrix factorization techniques for recommender systems"[3] provides the basic frame in matrix study and implementation.

3 DATA SET COLLECTION AND PROCESSING

The dataset is the sample dataset that is provided by myPersonality Project (http://mypersonality.org/wiki/)[2], which includes psychodemographic profiles of 110,728 Facebook users and their Facebook Likes records. There are 3 files in the dataset. The users.csv contains psychodemographic user profiles; the likes.csv contains anonymized IDs and names of 1,580,284 Facebook Likes; the users-likes.csv: contains the associations between users and their Likes, stored as user–Like pairs.

3.1 Data Exploration

As is shown in the Fig 1 and Fig 2, the distributions of the numeric variables are skewed. According to the Fig 3, the variable of "political" has a lot of missing value. In the research, the missing value is discarded and only records that include "political" are used be trained and predicted the feature of "political". All the other variables were not influenced by the missing value in the "political".

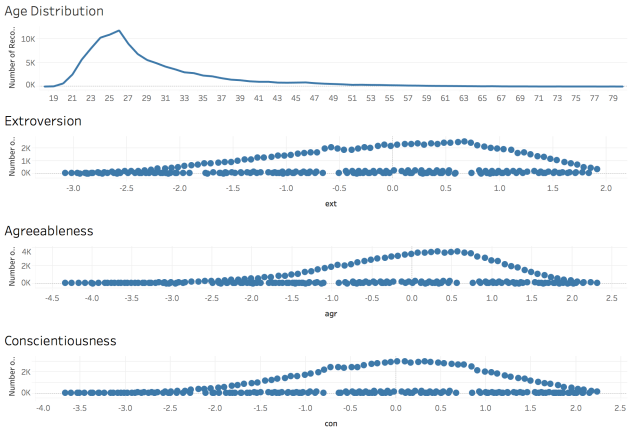

Figure 1: the distribution of numeric variables including age, extroversion, agreeableness and conscientiousness.

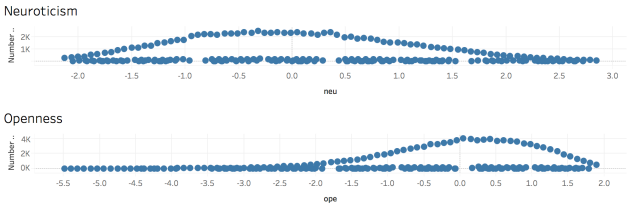

Figure 2: the distribution of numeric variables including neuroticism, and openness.

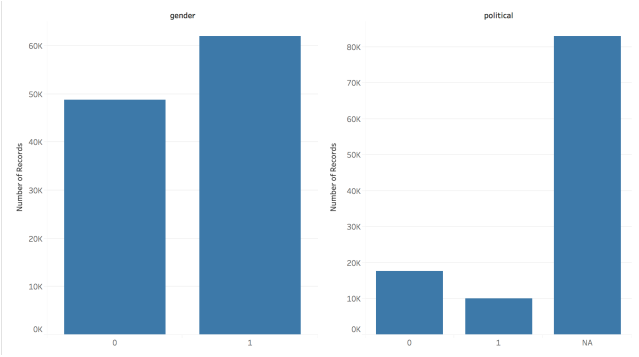

Figure 3: the distribution of categorical variables including gender, and political. For gender, "0" for male and "1" for female; for political, "0" for Democrat and "1" for Republican.

3.2 Data Set Collection and Processing

In this project, the user-Like pairs are transferred to a user-Like matrix. Because that the non-zero data is sparse in the matrix, the sparseMatrix was used to store the matrix. The dimension of the matrix is 110728 * 1580284, including 110728 users' Likes records of 1580284.

Because of the limited computing power, I used sample data set instead of the data set that was used in the original paper. The problem here is that the number of Likes per person on average is 170 in the original paper, while in the sample data set, the number of Likes per person on average is 95. To make the data set be closed to the original data, the users-likes matrix was trimmed to include users that has Likes more than 18. The threshold of 18 was chosen because of two reasons. First, the users that have too few Likes record are not useful for extracting patterns. Second, removing all the users that have Likes less than 18 makes the average number of Likes per person to be 170, which is the number in the original analysis.

4 METHODS

In this project of producing, all the method and parameters were tried according to the original work. Singular Values Decomposition (SVD), linear regression and logistic regression were performed. 10-ford cross-validation was also applied.

4.1 Singular Values Decomposition (SVD)

The Singular Values Decomposition (SVD) is used to reduce the dimension of Users-Likes matrix. In the original research, the SVD is performed with k = 100. To reproduce the result, SVD was also performed with k = 100 in this project. The package of "irlba" in R is the tool that was used to perform SVD to the sparse matrix.

4.2 Linear Regression and Logistic Regression

The same as the original work, the numeric variables (including age, openness, conscientiousness, extroversion, agreeableness and neuroticism) were predicted with linear regression model, while the dichotomous variables (including gender and political party) were predicted with logistic regression model. The function of "glm" is used in training the models.

5 RESULT

The Area Under Curve (AUC) and Pearson Correlation Coefficient are used to evaluate the accuracy of the predictions, just the same as the original work. In this project, the AUC of dichotomous variables were calculated with "ROCR" package. The result is shown in.Fig 4 However, The Pearson Correlation Coefficient of numeric features' prediction is pretty low, as is shown in Fig 5. The numeric variables' AUC can be calculated by the function of "multiclass.roc" in the "pROC" package, in which all the numeric values are treated as a class (shown in Fig 6). The pattern of the result is accordance with the result in the original research. The AUC of dichotomous features prediction is generally higher than the AUC of the numeric features.

The relation between analyzed Likes and the prediction accuracy is also studied. In the original paper, there is an obvious correlation between the accuracy and the number of analyzed Likes. However, by study the users who have 1 to 300 Likes in

records, the correlation was not found. The relationship between the prediction of accuracy and the number of analyzed Likes is shown in Fig 7.

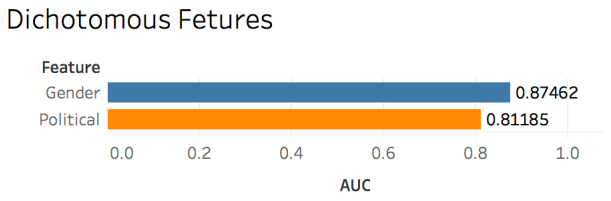

Figure 4: the AUC of prediction with logistic regression model for the categorical variables.

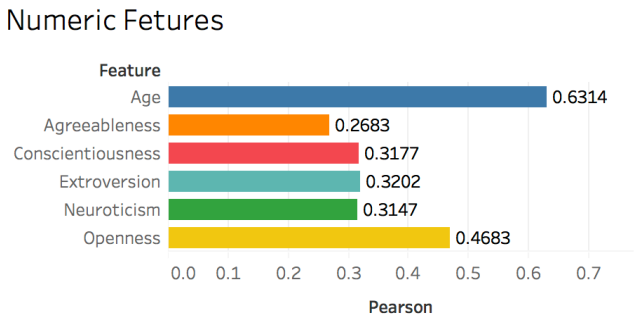

Figure 5: the Pearson Correlation Coefficient of prediction with linear regression model for the numeric variables.

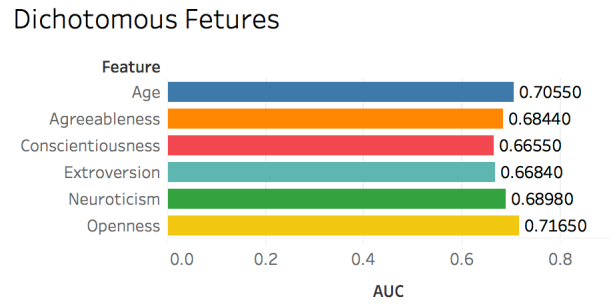

Figure 6: the multi-class AUC of prediction with loinear regression model for the numeric variables.

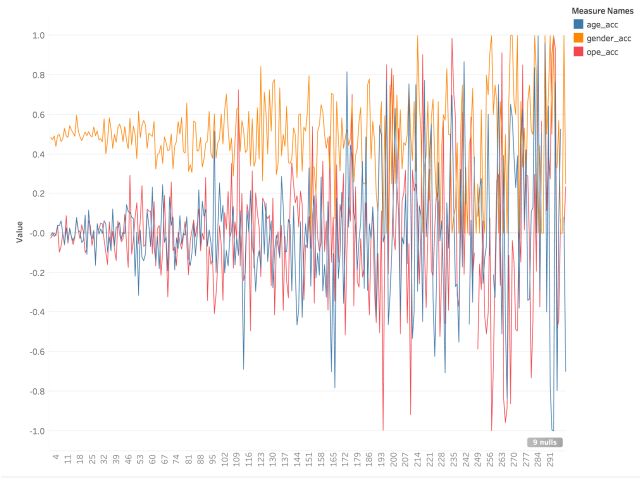

Figure 7: the accuracy when predicting age, gender and openness does not continuously increase with the number of Likes that analyzed.

6 COMPARE WITH THE ORIGINAL RESEARCH

6.1 Result

The most important argument in the original paper is that the personal attribute can be predicted from one's Like record. This argument was verified in this research. All the accuracy of the prediction of dichotomous features are higher than 80%. All the accuracy of the prediction of numeric features are higher than 60% (based on multi-class AUC).

The correlation between analyzed Likes and the prediction accuracy was not verified.

6.2 Data Set

The original research included far more Likes records and user information. The data set is larger than the data set in this research in number of features and number of observations. This element may cause the difference in the result. Although the data set in this research is trimmed to be as similar as the original data set, they are basically different any way.

6.3 Method

All the methods that were mentioned in the original paper were implemented in this research. The evaluation of accuracy is also based on original research..

7 CONCLUSIONS

The argument that "private traits and attributes are predictable from digital records of human behavior" is verified in this research. With the Likes records, the gender, political views, age, and personality can be predicted with 60%-80% accuracy. However, the relationship between the prediction accuracy and the number of Likes that is analyzed need further study.

Due to the time limitation and computing power limitations, the smaller data set was used in this research. The complete data set and all the other user information should be included in this reproducing research.

REFERENCES

[1] Kosinski, M., D. Stillwell, and T. Graepel. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." Proceedings of the National Academy of Sciences 110, no. 15 (2013): 5802-805. doi:10.1073/pnas.1218772110.

[2] "MyPersonality Project." Start [myPersonality Project]. Accessed April 24, 2018. http://mypersonality.org/wiki/doku.php.

[3] Mehta, Rachana, and Keyur Rana. "A Review on Matrix Factorization Techniques in Recommender Systems." 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), 2017. doi:10.1109/cscita.2017.8066567..

[4] Golub, G., and W. Kahan. "Calculating the Singular Values and Pseudo-Inverse of a Matrix." Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis 2, no. 2 (1965): 205-24. doi:10.1137/0702016.