UNIVERSITY OF MICHIGAN, CSE DEPARTMENT

Replication Project

Replication by Luke Brandl
Paper in question by Michal Kosinski, Yoram Bachrach, Pushmeet Kohli,
David Stillwell, and Thore Graepel
April 18, 2016

1 INTRODUCTION

In the paper "Manifestations of user personality in website choice and behaviour on online social networks", Michal Kosinski et al. suggested that there might be a strong link between a user's Facebook features and their personality. In the paper they used linear regression in an attempt to predict personality from these Facebook features, claiming that the results for linear regression were similar to other machine learning methods. The authors then claimed that this property supported their claim, noting that the ability to predict personality from the data is intrinsic to the data rather than a particular analysis method. As part of this argument, the paper stated:

> The results using these [non-linear regression] approaches are similar to the ones we obtained using linear regression, giving some support to our conjecture that the errors stem mostly from a high variance in personality traits even among people with similar Facebook profile features. (Kosinski et al. 2013a)

While obtaining similar results from other methods would certainly support their hypothesis, the authors never supported this claim with Pearson correlation coefficients for any algorithm other than linear regression. They even base their final conclusion for this study on this unsupported claim by stating:

> However, inspecting the relationship between log-transformed Facebook features and personality (that are relatively weak but predominantly linear), and the similar accuracy achieved using other prediction methods, indicate that the results presented in Table 9 (reproduced here as Table 2.2) do indeed capture the ability to predict personality using the Facebook profile features used in this study. (Kosinski et al. 2013a)

Overall all their argument and methodology is very solid, so this report is an attempt to rigorously verify the above claim in order to complete the argument presented in their paper. In particular, the paper presented strong evidence that Facebook features can predict extroversion, so this report will focus on this prediction. For a deeper discussion of this topic, see section 2.1

1.1 Background

The paper being examined, hereafter referred to as 'the paper', used on data from 354,000 Facebook users registered with the MyPersonality project (Kosinski et al. 2013a). The features used in their analysis included the number of friends, network density, grouped joined, things liked on Facebook, photos uploaded, statuses posted, photos tagged in, and events attended. Many of these features were only available for a small fraction of the data set, but all were available for at least 8977 users. The authors then tried to predict Costa and McCrae's Five Factor Model of personality, which was already known for each user from a questionnaire (Costa and McCrae 1992 and Goldberg et. al 2006). Among those five factors, their best model by far was extroversion, with a Pearson correlation coefficient of 0.31 out of a hypothetical 0.75 given the noise in the questionnaire.

1.2 Hypothesis

My original hypothesis was "Linear regression and other machine learning algorithms such as SVMs yield results within five percent of each other when

attempting to predict personality types based on myPersonality project data." Unfortunately I was not able to obtain the exact same 354,000 data points used in the original study, but Michal Kosinski advised me that data currently on MyPersonality.org is newer and a includes few more points, and that he expected the results to be very close[1]. As such, I decided instead to attempt to both reproduce the linear regression results and produce regression SVM results for the new data, as comparing SVM results for another data set to their regression results would yield very weak results. While this is not precisely confirming the unsubstantiated statement in their paper, if I can achieve similar results for what is essentially a new data set (see experimental procedure section) for both linear regression and SVM, it would strongly support the paper's hypothesis of Facebook features being able to predict personality type. As I will explore in detail in section 2.1, the data set I used is essentially a new data set. I also decided to only attempt to examine extroversion predictions, as their extroversion predictions were substantially better than any of the four other personality types. The other reasons extroversion was chosen over other attributes will also be discussed in section 2.1. As such, my hypothesis is:

> Linear regression and regression SVMs yield accuracy results within 0.05 of each other when attempting to predict extroversion based on the myPersonality project data, where accuracy is defined as the Pearson correlation between the predicted test extroversion labels and the 'true' test extroversion labels.

The labels for personality are not necessarily perfect, but the high IPIP scale reliabilities in the MyPersonality data suggest that the quality of those responses are at least as high as traditional personality test methods (Kosinski et al. 2013a). It would be fairly straightforward to expand this hypothesis to general Facebook data, but this is strictly a replication report so the hypothesis is somewhat focused.

2 Experimental Procedure

2.1 Obtaining and Preparing the Data

In order to obtain the data from mypersonality.org, I contacted Professor Michal Kosinski to have access to the necessary tables. I was able to obtain data

[1] From a series of emails beginning April 20th, 2016

for all of the features except photos uploaded, which I was unable to access for privacy reasons. While Professor Kosinski was not able to provide me with the exact data used in his study, he assured me that he expected the results to be very similar[2]. The database is quite large and contains dozens of files, the vast majority of which were not used in the paper in question. The personality data file had over five million entries while the data set used for the paper had only 354,000, so it was clear that some subset of user ids was taken. Starting with the several million user Facebook information and personality information files, I examined the size of the files containing other features. The semi-essential egocentric network data file contained around 80,000 users worth of data for the number of friends and network density, but it would reduce the data set to a much smaller size than the data set used in the paper. The 'month since joining Facebook' data set was much more promising, as it contained around 722,000 records.

While not actually a feature used in the study directly, the information in this last data set is critical because many of the Facebook features used by the paper's model are expected to increase over time, so those features were divided by the time the user had been using Facebook. As five of the seven features I used depended on this time, all user ids which could not be found in this date joined data set were removed. The number of users left who had personality and time since first using Facebook was 374,965, suggesting that this is likely the newer iteration of the 354,000 users used in the initial paper with a few new users added over the years. These two attributes are absolutely essential for the study, as one contains the labels and the other contains the information needed to normalize five of the eight features, so the above statement is fairly well supported. I completed the data set by looking up Facebook and network features for those 374,965 users. Given that only these users have all of the information critical for the study from the paper, I believe that this data set was reasonable to use for replication. A statistical summary of the data obtained as well as the data from the original is contained in Table 2.1.

[2]From a series of emails beginning April 20th, 2016

Label	Details	n	μ	$\hat{n}$	$\hat{\mu}$
Friends	Number of friends	62,780	295	31486	337
Network Density	Density of friend network	28,835	0.03	31486	0.045
Groups	Groups joined	97,622	15	63718	38
Likes	Things liked	109,837	82	67706	171.5
Photos	Photos uploaded	110,540	10	349675	24
Statuses	Statuses posted	71,912	99	50577	143
Photo Tags	Photos tagged in	315,169	20	-	-
Events	Events attended	8977	10	10212	32

Table 2.1: This table enumerates the features of the data set, as well as the number of points containing information for each feature n and the mean for the feature μ. The last two columns are the same values, but for the the new data set of 374965 users.

Interestingly, the number users with a non-missing attribute decreased for every attribute except for network density, despite there being over twenty thousand more users in my data set. I suspect this may be due to Facebook's efforts to increase users' privacy awareness over recent years, but have not investigated that claim. All averages have gone up, which supports the idea that many users have been updated since the original experiment, as those users would likely have more photos, likes, etc. now than they did several years ago when the data for the initial paper was drawn. This also reinforces the need to regularize most of the parameters to be regularized by time since joining Facebook. The number of points containing each of the six features used in the paper to predict extroversion actually increased slightly (16,900 to 17,135), likely indicating that many users still gave nearly all of their Facebook information to the project. Overall my data set is not identical to the original, but I believe that it is appropriate to apply the same methods used in the paper to extract personality type from this data. In fact because this data set can almost be considered an entirely new one, a similar accuracy would strongly support the both link between extroversion and Facebook data and the model suggested to predict extroversion from Facebook data.

The model for predicting extroversion is the most accurate personality prediction model presented in the paper, as expressed by Table 2.2, a table drawn from the paper. The model for extroversion also model included only Friends, Groups, Likes, Network Density, Photo Tags, Statuses as input features. The

Trait	Accuracy (r)	n	Features used in the prediction
Openness	.11	18,720	Friends, Groups, Likes, Network Density, Photo Tags, Photos
Conscientiousness	.16	18,720	Friends, Groups, Likes, Network Density, Photo Tags, Photos
Extroversion	.31	16,900	Friends, Groups, Likes, Network Density, Photo Tags, Statuses
Agreeableness	.05	45,565	Friends, Likes, Photo Tags
Neuroticism	.23	9,515	Friends, Likes, Photos, Statuses

Table 2.2: A table from Kosinski et. al (2013a) which showed the accuracy expressed by Pearson correlation coefficient, as well as the features used in their linear regression model

only feature I am missing is photos uploaded (here just photos) due to privacy reasons, which was not used to predict extroversion. It also does not use Events, which is the most sparse feature in both the original data set and my new data set, which allows for many more points with all features present. The only traits that can be predicted using their models are extroversion and agreeableness, but their model for agreeableness had very weak performance. As such, I decided that focusing my efforts on extroversion would be most beneficial for this replication. I also believed that rather than optimizing the feature list on my own, using the features provided by their model would be most accurate way to reproduce the paper's results, as it would provide valuable insight whether their features were overfit to their specific data set, or actually indicating and underlying relation between Facebook use and extroversion.

I also examined other regression methods mentioned by the authors such as decision stumps, but the exact methods the author used to convert typically classification based methods into regression models were not specified in paper. I spent some time attempting to research what methods they might have used, but in the end I decided that two types of regression SVM kernels (for details see Bishop 2006) would be enough to show that the results shown in the paper were not specific to linear regression. I could not find any existing Matlab libraries for the methods they used, so I decided it would not be worth the time to implement a method they possibly did not use in the form I would implement it.

I explored state of the art methods of imputation for my extremely sparse data set, but in the end decided to just use the 17,135 data points which have all six features present. Because the features were so sparse with many being in less than ten percent of the points, I felt that predicting missing values would make the already high variance of personality prediction even worse. The paper did not specify exactly which points they used, but for their extroversion prediction they used 16,900 data points, so I suspect that they also took only the complete features from their slightly smaller data set. Regardless of their methods, I believe that using their features for 17,135 complete points should yield similar accuracy to their predictions for both linear regression and SVMs. I used the same preprocessing steps as defined in the paper of dividing friends, groups, likes, photo tags, and statuses by the number of months the user has used Facebook. I then log transformed the data as in the paper.

2.2 The Experiment

For the experiment I used 17135 data points, and tried to follow the procedure identified in the paper as closely as possible. I used ten fold cross validation as in the original paper, so the r values presented are the average of the ten folds. The Support Vector Machine regression experiments used $\epsilon = 0.1$, and $C = 1$, with the polynomial kernel one using maximum dimension 3. For details on the regression SVM methods used, see (Bishop, 2006). The results observed were:

Data Set	Model	Accuracy(r)	n
Original	MV Linear Regression	0.31	16900
New	MV Linear Regression	0.27	17135
New	Regression SVM RBF Kernel	0.29	17135
New	Regression SVM Poly. Kernel	0.27	17135

Table 2.3: This table shows the accuracy of extroversion predictions, modeled by Pearson correlation coefficient between the predictions. The 'Original' and 'New' data sets refer to the original data set analyzed in the paper, and the new data set used for my results.

3 Discussion and Conclusion

The accuracy of linear regression and RSVM are in fact very similar to each other, being well within the 0.05 needed to satisfy the hypothesis. In fact, even the original linear regression accuracy on a much older version of the data set used in the original paper has similar accuracy to the new RSVM accuracy. These results strongly support the idea that extroversion can be predicted to some extent by examining Facebook features. Seeing as extroversion was by far the trait most strongly expressed by Facebook features (see Table 2.2), the confirmation of my hypothesis does support their most firm claim that extroversion can be predicted by aggregate Facebook data. Obviously my results do not make as strong of a statement about the features I did not examine, but with a missing feature I could not test their model on any trait but agreeableness, but their accuracy was so low for agreeableness it was not worth testing.

Another characteristic observed is a lower accuracy altogether across the new data set. This is to be expected somewhat, as the authors carefully tuned their features in order to maximize performance, where I only recreated the model they optimized for. This actually supports their chosen features capturing more than just over fitting the data. I spent some time trying to optimize SVM parameters, but ended up finding fairly consistent results over my parameter sweep. It is very plausible that if I were to spend more time optimizing regression features and parameters that the newer data set accuracy can reach the 0.31 achieved in the paper.

Overall, my replication efforts were successful in that I verified the authors' claim that linear regression and SVMs yielded similar results for predicting extroversion. While the accuracies achieved in this paper were slightly worse than in (Kosinski et al. 2013b), it is still useful to know that aggregated Facebook likes can also be used to indicate extroversion. My replication efforts also show that their model was not overfitted to their data set, and likely would be able to be applied directly to other facebook data sets directly.

4 BIBLIOGRAPHY

Bishop, C. M., and SpringerLink (2006). Pattern recognition and machine learning (Vol. 4). New York: Springer

Costa, P. T. Jr., and McCrae, R. R. (1992). Neo personality inventory-revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual. In Psychological assessment resources, Odessa, FL.

Goldberg L.R., et al.Five Factor Model of Personality (2006). The international personality item pool and the future of public-domain personality measures. J Res Pers 40:84-96.

Kosinski, M., Matz, S., Gosling, S., Popov, V. and Stillwell, D. (2015) Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. American Psychologist.

Kosinski M., Bachrach Y., Graepel T., Kohli P., Stillwell, D (2013a). Manifestations Of User Personality In Website Choice And Behaviour On Online Social Networks. Machine Learning Journal (MLJ), 95, 357-380.

Kosinski, M., Stillwell, D., and Graepel, T. (2013b). Private traits and attributes are predictable from digital records of human behavior. In Proceedings of the National Academy of Sciences.