Reviewer's report

Title: Applying Machine Learning to Predict Future Adherence to Physical Activity Programs

Version: 0 Date: 23 Jun 2019

Reviewer: Chen Liang

Reviewer's report:

This paper reported the experiments using machine-learning classification to predict exercise relapse followed by a downstream analysis (simulation) of the classified data with the goal of identifying the best indicators for effective prediction of users' behavior patterns and controlled intervention strategies. In this design, authors proposed Discontinuation Prediction Score (DiPS) as an artificial indicator shown superior effectiveness. The idea of DiPS is new in the field. The paper is strong in the delicate design of the statistical process and experiment. However, the presentation of the logic flow, procedures, and technical details is not clear on multiple occasions. Please find my comments below for authors' consideration.

Major concerns:

1. Page 8, Pre-processing. The rationale of augmenting the training data needs to be clarified before introducing the technical details.

2. Page 9. I tend to think that the rationale of adopting logistic regression and SVM is not strong. Both of them are well established, true. But I wouldn't fully agree with the statement that SVM has strong interpretability unless authors could provide a comprehensive justification. In addition, it is important to explain why these two methods are a good fit for the present study (e.g., dataset, features, performance, published benchmark work, etc.), preferably with pertinent references. On page 13, authors stated that tree-based algorithms "may also be applied to this problem and generate comparative results". That was an abrupt sentence. I had a difficult time understanding what authors were trying to express. And why three-bases algorithms were not used in the experiment.

3. It is not clear at the first place what the classifier tried to classify and what role the DiPS plays in the classification until I read the Results and Discussion. It is suggested to provide an overview/workflow of the machine learning experiments at the beginning of the Methods.

4. Is there an algorithm considered as the benchmark classifier? Or is there a published work serving as the benchmark performance? If the present work/data-set is unique or the experiments are exploratory, please make a clarification.
5, The technical details for the evaluation of machine learning experiments is needed, e.g., 10-fold or other design, number of trials.

6, Page 12. Please clarify how you selected a sequence of values for p*, step*, and DiPS*. Page 13-14, where does the threshold of 0.5 come from?

7, Page 13. Authors stated that Lasso, Ridge, and Elastic methods were tested but did not outperform. If this is a part of the study design, all details should be reported. If they are side experiments - and I would like to ask why - details could be reported as supplements. Also related to Major Concern 2, I think it's important to have a scientific rationale for the selection of algorithms in the machine learning experiments.

8, Page 15, Model Interpretation. Explain "coefficient". Authors also mentioned the "coefficient" for SVM. But Table 4 only shows coefficients for logistic regression. What is the coefficient from SVM?

Minor concerns:

1, Page 4, the last paragraph, "Despite the wide adoption……" The font seems not consistent with the rest part of the manuscript.

2, Page 8, "The features on last week behavior capture the immediate past performance, which has been shown to be most predictive for the immediate future." Is this conclusion drawn from any analysis during the feature engineering or a published work? References are needed if later one.

3, Page 8, "i.e., the complete set of features for week 3, 4, and 5 respectively……". The definition of DiPS (page 9) should be provided before using it in the feature pre-processing.

4, Consider it a suggestion. The mathematical mechanism for logistic regression and SVM could be simplified unless revised algorithms were proposed in the study. These are well-established methods that the informatics communities are familiar to.

5, The use of verb tenses is not consistent. It's difficult to understand what is done and what is not.
6, Page 14. The comparison between the three groups is descriptive with some inaccurate conclusions. For example, "The PLUS group has the highest test AUC for most of the weeks during the maintenance period. The test AUC on the REGULAR group is between those of the CONTROL group and the PLUS group." Use solid data/statistics to specify "most" and "between".

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Yes

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

Yes

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published
Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal.