RESPONSES TO REFEREE COMMENTS FOR HICSS 2019 ARTICLE ID 46 IN THE MACHINE LEARNING AND PREDICTIVE ANALYTICS IN ACCOUNTING, FINANCE AND MANAGEMENT SECTION

1. Introduction

The following are comments related to revisions of the HICSS 2019 submission *An Explicative and Predictive Study of Employee Attrition using Tree-based Models*

We are grateful to the chair and three referees for evaluating our manuscript and for providing us with detailed comments and suggestions for improvement of both the language and content of this article. In accordance with these comments and suggestions, we have carefully revised the paper. All issues raised in the referee reports have been addressed thoroughly.

In what follows, we detail the changes made according to the referees' recommendations. For convenience, important excerpts from the referee reports are printed in blue, whereas our revision related statements are printed in black. In the revised manuscript, the changes are printed in blue.

2. Response to Referee 1

(1) As most of the research paper includes the summary of some of the most relevant researches to motivate the applicability of the research, it would have been better if the paper included some similar works on the topic. It would have given good impact on readers about why is this paper different from others.

(2) In the modeling, Random forest is used to train model with original features but then Light GBT is used when training with additional features. It would have been better if Random forest was included in the later approach as well for clear comparison.

(3) Also, you mention many several other models were trained for binary classification, but no results of such models are reported. The summary of those model results if presented would have helped in creating more impact in the results presented.

(4) Few changes - You have no indentation on the first paragraph at every section - check the formatting.

(5) In Table 1.- an extra caption may not be necessary as you already have table heading and contents of the table described.

3. Response to Referee 2

(1) An important part that is not presented in sufficient detail would be to identify and discuss related contributions from the literature. The authors identify three articles, those should be discussed in more detail; additionally the authors should present what are the most important other application areas of machine learning in human resources management, in particular on the use of tree-based models.

(2) After the literature review and before data description and summary, the authors should include a methodology section, summarizing the basics of predictive modelling, a brief description of the utilized models (this is done now in different places in the paper, but mainly in 4.1-4.3), and the evaluation metrics for binary classification problems. While the limitation on the length of the paper does not allow for a detailed discussion on each model, the author should at least mention the name of each model they used, as in the article only few are named and then the authors mention that they used many more.

(3) The data description and descriptive analysis is well-presented and offers some interesting findings, however, the lengthy presentation of this basic analysis did not leave space for the man model building which is presented now in essentially one page if we exclude the theoretical background on the models. I would recommend the authors to shorten Section 3: keeping all the essential results but only discuss them briefly, most of the figures present specific details that are simply repeated in textual form in the manuscript.

(4) Regarding the models, as the authors aim to predict a binary value, applying simple linear regression is not the most correct way to approach this problem as many underlying assumptions tend to be violated with this type of data. I would suggest to exclude that from the analysis unless there is some specific points to make about it (or it is important to compare to the previous study mentioned several times in the paper), and use logistic regression as the baseline model to compare other models to.

(5) As mentioned above, specify all the methods that were used in the experiments. There is no mention on the implementation of the experiments, what libraries are used etc., this should be included. In the model performance presentation, additional models tested could be included in the figures, and additionally to the ROC curve, the AUC values could be presented in a table as a more quantitative comparison, for all the models, not only for the ones depicted in the figures at the current form.

4. Response to Referee 3

(1) Excluding PCA features, the best ROC (73%) is obtained by the random forest model. According to this model, what is the ranking of features in terms of importances?

(2) In the paper, it is not clear whether PCA features are first obtained, and then the dataset is split into training and test set; Or the dataset is first split into training and test set, and then the PCA features are obtained based on the training set. It must be clarified. Intuitively, it seems that PCA features are obtained before splitting the dataset. If so, the information related to the PCA analysis is embedded into the test set. In this case, a better performance by the light GBT can not be generalized for unseen data in future.

(3) When applying PCA analysis, it is assumed that the range of features in the training data and future unseen data is the same. In the problem considered, it does not seem that in future, we see features' value within the range of the training dataset. Therefore, adding PCA-based features may not be useful.

(4) The dataset is imbalanced in terms of number of labels. In terms of Precision Recall area under curve, how was the performance of the models? When splitting the dataset into training and test sets, how similar (dissimilar) are features' distribution? If the distribution of features are dissimilar, there might be some clusters in the dataset. In this case, some clustering analysis can help to divide the dataset into some clusters with similar samples. Then a model for each cluster can be built separately. This helps to achieve higher ROC for each cluster.