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 minitrack 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.

   We agree with the referee that a current and more thorough literature review was missing from the initial draft of this article. We have added an additional paragraph in the Introduction section that describes similar articles where employee attrition is examined from a predictive modeling perspective. In addition, in the subsequent paragraph, we described several ways in which our dataset and methods differ and extend upon this prior work.

(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.

   We added in the random forest model to the full feature example in the updated version of this article.

(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.

   We added the names and associated area under the ROC curve values for five additional strongly performing models we considered in the full feature case in the last paragraph of the Model Performance Comparison section.

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

   We initially utilized the HICSS latex .cls file in the preparation of this draft which does not appear to indent the first paragraph in any section, subsection, etc. In the most recently revision, we manually added indentations to these paragraphs.

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

We have now combined the caption and table heading in Table 1.

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.

We agree with the referee that a current and more thorough literature review was missing from the initial draft of this article. We have added an additional paragraph in the Introduction section that describes similar articles where employee attrition is examined from a predictive modeling perspective. In particular, we have focused on reviewing recent literature concerned with the application of tree based models to employee attrition prediction. In addition, in the subsequent paragraph, we described several ways in which our dataset and methods differ and extend upon this prior work.

(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.

We added an additional paragraph in the location recommended by the referee that summarizes the basics of predictive modelling as well as the core ideas behind the main models considered in this article as well as associated binary classification evaluation metrics. We were brief in this discussion in order to comply with page length article limitations. In addition, we mentioned the names of several additional models that were considered during the evaluation process but would not discussed in Section 4 in the introduction.

(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.

We shortened Section 3 by removing several sentences that described conclusions that one can directly draw from the Figures and retained text describing unique aspects of the attrition dataset under consideration. To the referee's second point, we would have liked to provide further explanations of technical details related to construction of the tree-based models considered in this article. However due to space limitations, and since this material may be readily found in standard references, we decided to focus text on describing specific summary information of the dataset.

(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.

We agree with the referee that linear regression based classification model is not the most appropriate model for this employee attrition prediction application. It is included in order to compare against the methods developed the Smart and Chamberlain reference as one of the main aims of the article is to extend and improve upon their results. To the referee's point, we added additional text to indicate that the logistic regression model should be viewed as the baseline model.

(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.

We added a paragraph in the introduction to Section 4 that describes the Python packages that were utilized for model implementation, comparison, and visualization. We added AUC results for five additional strongly performing classification models in the full feature case. We agree with the referee that a more comprehensive table summarizing the quantitative comparison between the models considered would be ideal; however, due to page limitations, we only summarized the models and performance metrics we believe to be most valuable.

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?
We added a discussion of feature importance based upon permutation testing for the light gradient boosted tree model in the full feature model section.

(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.

Although the full data set was used to construct the PCA features in section 3.4, we only considered training set data in the predictive model section 4.4. Specifically, a PCA was performed only on the training set and associated features rely only upon training data. We added a clarification note state the above in the article.

(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.

The PCA features are based on quantile normalized ratings values. The range of possible values of the PCA features is the same for both the training and test sets. In the specific test set that we considered, the range of actual PCA features was a subset of that of the training set. Overall, we found that they did add some additional predictive power to the tree based models being considered, although they were far from being the more important features from a performance perspective.

(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.