Review of: "Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis"

Nagendra Panini Challa[1]

1 Vellore Institute of Technology

Potential competing interests: No potential competing interests to declare.

Review comments:

1. This work demonstrates the utilization of the Synthetic Minority over Sampling Technique (SMOTE) on a dataset related to bank turnover. The SMOTE technique was utilised to tackle the issue of data imbalance, while the Genetic technique (GA) was implemented to choose the most informative characteristics from the original dataset. The evaluation of the selective features was conducted using four distinct classification algorithms: Random Forest (RF), K-Nearest Neighbour (KNN), Artificial Neural Network (ANN), and AdaBoost techniques. The KNN model exhibited exceptional performance in terms of accuracy (96%), precision (96%), and F-measure (96%) when compared to other models.

The authors of the study utilised traditional classification approaches such as SMOTE, GA, KNN, ANN, RF, and AdaBoost for churn prediction. The choice to utilise these established methods may be attributed to their proven effectiveness and widespread usage in the field. Lack of novelty.

2. There are numerous balanced datasets available in repositories. What is the purpose or rationale behind imbalancing? The current dataset being utilised is neither real-time nor proprietary.

3. The introduction section needs some elaboration.

4. A survey needs to be tabulated.

5. The Genetic Algorithm (GA) has been utilised for the past 20 years. What is the reason for selecting this algorithm, and what is the size of the reduced dataset (considering the original dataset with 12 features)?

6. The research methodology framework is insufficient and requires enhancement in its architecture.

7. The feature selection Genetic Algorithm (GA) will be used to address the bank churn problem.

8. The equation numbers are not specified.

9. The dataset obtained from Kaggle's repository is nearly evenly distributed. According to Table 2, the Random Forest (RF) algorithm achieved the highest accuracy of 1.00. However, you mentioned that the K-Nearest Neighbours (KNN) algorithm is preferable. Could you please explain why?

10. The reason for explaining the confusion matrix (4.4.1) after the results section is to provide a clear and concise

understanding of the performance evaluation of the model.

11. Table 6 displays the comparative research. Was the same dataset used for all the studies conducted in this research?

12. Include the name of the suggested model (KNN) in the conclusion section.

13. How do authors identify gaps in earlier work using only six references? A minimum of 20–25 references should be included.

14. Significantly, the study did not examine the attributes of expected customer attrition, despite the fact that such knowledge might be extremely important for organizations in determining whether to keep or let go of specific clients. Hence, further studies will focus on analysing customer turnover characteristics, as they have the potential to provide more significant long-term benefits to the organization. Please provide a rationale or logical reasoning to support or validate this statement.

15. Additional clarification is required in the preprocessing section. Authors are currently not focusing on this issue.