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

Shourjya Sanyal[1]

1 Unum Group (United States)

Potential competing interests: No potential competing interests to declare.

The study provides valuable insights into the feasibility of machine learning models to churn production in finance. Particularly, the work compares the accuracy of four types of models - Random Forest (RF), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Adaboost algorithms.

The interesting bit is the result shows almost human-like accuracy with KNN models.

Scope of improvement -

1. Abstract -

1A. Banking is a huge enterprise, so add further details about the "bank churn dataset." What product? What customer?

1B. Neighbour typo!

2. Introduction -

2A. "Consequently, this presents a considerable challenge for newer service providers to devise innovative strategies to meet and exceed customer expectations." Same challenges for new and old! So remove the word "new."

2B. In the Introduction, can you rephrase the last paragraph to emphasise why your approach is better? "In this study, we deploy genetic algorithms, since they can ..."

3. Review of Related Works -

3A. Can you add a paragraph or two about how churn prediction was historically done without ML? No need for describing individual papers, but broad stroke approaches and referencing to multiple papers for each approach.

3B. "The Kaggle churn modelling dataset served as the basis for analysis" - Add reference.

3C. Domingos et al., (2021) - Comment - Churn is a difficult space as different models might be better for different types of churn, high churn, medium churn, low churn.

So running your models for multiple use cases in each of these categories is required to get a global perspective.

3D. For the studies described in "Review of Related Works," could you create a table showing types of models used and

the input data types used as inputs? It will give readers an easy way to understand the current state of the art.

4. Research Methodology -

4A. Rephrase "For the evaluation of bank customer churns, it is essential to devise a methodology. Hence, the proposed research methodology consists of three main phases:" to "For the evaluation of bank customer churns, our research methodology consists of three main phases:"

4B. Dataset Descriptions -
Add a summary table describing the data set - How many data points, how many unique individuals, etc.
For each of the variables, can you give a range and type? Say, Credit Score is 0 to 10. Also, compute the average value of each category (mean and median). This will help readers view the data sets.

Not sure how churn is quantified.

4C. In Fig. 1, it is unclear how "Check Data Imbalance" decides to apply or not apply SMOTE.

5. Results and Discussion -

5A. Performance Evaluation Metric - The formulas did not render correctly.

5C Figure 8. Comparison of Accuracy results of Classification with and without SMOTE - The Y-Axis did not render correctly, 0.98 not 0,98.

6. Additional Work - Could you break the system into 25%, 50%, and 75% and show if the accuracy matrix still holds? If you see the same trends, it will show that KNN with SMOTE is truly best in class.