Assignment 2 (Group)

In many real-world applications such as fraud detection, medical diagnosis, and predictive maintenance, the datasets involved are highly imbalanced, where one class (often the event of interest) occurs far less frequently than the other. Traditional machine learning algorithms tend to be biased toward the majority class, resulting in misleading performance metrics such as high accuracy but poor detection of rare but critical events.

This assignment challenges you to work with a given dataset containing a significantly imbalanced target variable. Your task is to explore, model, and critically evaluate various strategies for handling class imbalance in a predictive modeling context. The assignment will not only test your technical ability to apply techniques like SMOTE, undersampling, and class-weight tuning, but also your ability to interpret the results, justify methodological choices, and make deployment-oriented recommendations.

You are expected to go beyond textbook solutions and demonstrate understanding through experimentation, implementation, and reflection. Remember, accuracy is not always the best measure. Your insight into the problem's context and model trade-offs will be key.

Here are the detailed instructions:

Part A: Exploratory Analysis and Class Distribution

1. Briefly describe the dataset.
2. Visualize class distribution.
3. Discuss the implications of class imbalance on model performance.

Part B: Baseline Model Without Resampling

1. Train a logistic regression and decision tree model without addressing imbalance.
2. Report accuracy, precision, recall, F1-score, and AUC.
3. Discuss why the accuracy might be misleading.

Part C: Addressing Imbalance

Apply at least 3 different methods to handle imbalance such as:

- Random undersampling / oversampling
- SMOTE or ADASYN
- Class-weight adjustment
- Cost-sensitive learning
- Threshold moving
- Balance bagging

For each method:

1. Explain the method briefly
2. Retrain the best model from Part B.

3. Report all metrics again.
4. Discuss the differences/improvements.

Based on all findings, recommend a final model for deployment.

Submission Format:

- Final report in PowerPoint slides (max 10 pages): concise summary, model comparison table, plots and final recommendation. [10 marks]
- Prepare the presentation video (less than 10 minutes). [20 marks]