Selected Research Approach: Quantitative

To examine the impact of data balancing approaches on model performance in imbalanced classification problems, a quantitative approach has been used for this research project. The quantitative approach involves analyzing and measuring various performance metrics to evaluate the impact of different data balancing techniques.

Research Plan

The research plan shows the sequential steps involved in conducting the study. It outlines the key stages, tasks, and the logical flow of the research process.

Step 1: Literature Review: Conduct an extensive review of existing literature on class imbalance solutions, data balancing techniques, and their impact on model performance.

Step 2: Dataset Selection: Identify suitable datasets that exhibit class imbalance and are relevant to the research objectives. The "data" subdirectory in the GitHub repository will contain dataset links.

Step 3: Data Preprocessing: Perform necessary data preprocessing tasks such as handling missing values, removing outliers, and encoding categorical variables.

Step 4: Baseline Model Building: Build initial models using the imbalanced dataset without applying any data balancing techniques. Evaluate their performance using appropriate evaluation metrics.

Step 5: Data Balancing Techniques: Apply various data balancing techniques, including oversampling (e.g., SMOTE, ADASYN) and undersampling (e.g., RandomUnderSampler, NearMiss), to address the class imbalance issue in the dataset.

Step 6: Model Training and Evaluation: Train and evaluate models using the balanced datasets obtained after applying different data balancing techniques. Compare their performance with the baseline models.

Step 7: Performance Analysis: Analyze and compare the performance metrics (e.g., accuracy, precision, recall, F1-score) of different models to assess the effectiveness of the data balancing techniques.

Step 8: Results Interpretation: Interpret the results and draw conclusions regarding the impact of data balancing techniques on model performance in imbalanced classification tasks.

Data Collection Strategy

- The research project will utilize publicly available datasets that exhibit class imbalance and are relevant to the research objectives.

- At least 1 or 2 datasets will be used for each type of imbalance (Mild, Moderate, Extreme)
- Dataset links and details will be added to the "data" folder of the GitHub repository.
- The selected datasets will cover various domains, ensuring a diverse representation of imbalanced classification tasks.
- The datasets will be sourced from reputable repositories and publications, ensuring data quality and reliability.

Next Steps:

- The next phase of the research project involves data preprocessing, building baseline models, and applying data balancing techniques to evaluate their impact on model performance.
- The subsequent steps will include model training, evaluation, and performance analysis to draw meaningful conclusions regarding the effectiveness of different data balancing techniques.