Objective Document: Addressing Class Imbalance in Binary Classification

Objective

The purpose of this task is to address the challenges posed by class imbalance in binary classification problems. Using the IDA2016 Challenge dataset, we aim to build and evaluate classification models that effectively handle a severe class imbalance (class ratio of 1:59). The ultimate goal is to improve the macro-average F1 score of classifiers through various techniques.

Dataset Overview

- **Dataset**: IDA2016 Challenge dataset
- **Problem Type**: Binary classification ($$y = \{\text{'pos', 'neg'}\}$$)
- **Features**: 170
- **Data Points**: 60,000
- **Class Ratio**: 1 positive sample for every 59 negative samples
- **Files Provided**:
    - Training file: `aps_failure_training_set.csv`
    - Testing file: `aps_failure_test_set.csv`

For this task, only the training file will be used for experimentation.

Tasks

Task 1: Baseline Classifier Development

1. **Data Partitioning**:
    - Split the training dataset (`aps_failure_training_set.csv`) into train and test partitions.
2. **Model Development**:
    - Build baseline classifiers using the following algorithms:
        - Support Vector Classifier (SVC)
        - Logistic Regression (LogReg)
        - Decision Tree (DT)
    - Perform hyperparameter tuning using `GridSearchCV` with 5-fold cross-validation:
        - SVC: Tune parameters such as kernel type and kernel-specific parameters.
        - LogReg: Tune regularization type (L1/L2) and regularization parameters.
        - DT: Tune tree depth and leaf size.
3. **Evaluation**:
    - Train models on the training partition using the best hyperparameters identified.
    - Report performance metrics on both train and test partitions.

Task 2: Addressing Class Imbalance

To improve classification performance, apply the following techniques across all three classifier families:

1. **Resampling Techniques**:
   - Undersample the majority class.
   - Oversample the minority class.
2. **Class Weight Adjustment**:
   - Use weights inversely proportional to class frequencies (`class_weight`).
3. **Sample Weight Adjustment**:
   - Assign penalties for misclassifications based on the data point's class using `sample_weights`.
4. **Creative Solutions**:
   - Explore and implement additional innovative methods to handle class imbalance.

Goal:

The modified classifiers should demonstrate improved macro-average F1 scores compared to baseline classifiers.

Additional Notes

- Preprocess the dataset to ensure it is suitable for building classifiers.
- Focus on improving model performance while addressing the imbalance challenge effectively.