Overview of Project Repository

This project repository is designed to address a multi-class classification and regression problem using advanced machine learning techniques. The solutions are tailored to tackle real-world challenges such as:

- Class Imbalance: Managing unequal distribution of classes in the dataset.
- Overfitting: Ensuring the models generalize effectively to unseen data.
- Hyperparameter Optimization: Fine-tuning parameters to maximize model performance.

The repository adopts a well-structured approach, encompassing data preprocessing, model development, and comprehensive performance evaluation.

Tasks

1. Classification Task

The classification task focuses on categorizing social media accounts into 10 distinct categories based on their usernames. These categories include influencers, organizations, and other predefined groups.

Challenges

- Class Imbalance: Some categories have significantly fewer examples, leading to skewed data distribution.

Performance Targets

- Classification Accuracy: Achieve an accuracy of over 90%.
- F1-Weighted Score: Achieve a score of over 90%, ensuring balanced performance across all classes, particularly minority ones.

Approach

- Gradient Boosting Classifier: Utilized algorithms such as XGBoost, LightGBM, or CatBoost for their effectiveness in handling class imbalance.
- Solutions for Class Imbalance:
  - Oversampling minority classes using SMOTE (Synthetic Minority Oversampling Technique).
  - Applying weighted loss functions to account for the imbalanced class distributions.

Expected Outcome

The classification model is expected to categorize social media accounts into 10 distinct categories with high accuracy and balanced performance. This capability is particularly valuable for applications such as social media analytics and user behavior analysis.

2. Regression Task

The regression task aims to predict the number of likes a social media post will receive based on attributes such as content type, engagement metrics, and hashtag usage.

Challenges

- Large Fluctuations: Social media posts exhibit significant variations in the number of likes, resulting in high variability in predictions.

Evaluation Metric

- Logarithmic Mean Squared Error (Log-MSE): Prioritized as it penalizes large errors while being tolerant of small deviations.

Approach

- Gradient Boosting Regressor: Developed a robust regression model to handle the variability in the target variable.
- Logarithmic Transformation: Applied a logarithmic transformation to the number of likes to reduce variance and improve prediction accuracy.
- Feature Engineering: Extracted meaningful features such as engagement rates and time-of-post effects.

Expected Outcome

The regression model is expected to accurately predict the number of likes for social media posts, enabling businesses and content creators to better understand and forecast the performance of their content.

Repository Configuration

The repository is organized into three primary components:

1. Data Preprocessing Scripts

- Cleaning: Addressing missing values and outliers.
- Feature Extraction: Deriving meaningful insights from usernames and post attributes.
- Data Enrichment: Enhancing the dataset by incorporating external information.

2. Gradient Boosting Models

- Classification: Using Gradient Boosting algorithms (e.g., XGBoost, LightGBM, CatBoost) for multi-class classification.
- Regression: Utilizing Gradient Boosting regressors for predicting the number of likes.

- Hyperparameter Optimization: Leveraging techniques like grid search and early stopping to enhance model performance.

3. Evaluation and Documentation

- Performance Metrics:
    - For Classification: Accuracy, F1-score, and confusion matrix.
    - For Regression: Log-MSE, Mean Absolute Error (MAE), and R^2.
- Documentation of Findings:
    - Strengths and weaknesses of the models.
    - Observations on data challenges and model behavior.
    - Suggestions for future improvements.

Performance Targets

1. Classification Task

- Accuracy: Over 87%.
- F1-Weighted Score: Over 88%, ensuring balanced performance across all classes, including minority ones.

2. Regression Task

- Log-MSE: Minimize Logarithmic Mean Squared Error.
- Mean Absolute Error (MAE): Ensure prediction errors remain low.
- R^2: Explain a significant portion of the variance in the target variable.

Conclusion and Recommendations

This project repository offers comprehensive solutions to two key problems in social media analytics:

1. Multi-Class Classification: Accurate categorization of social media accounts.
2. Regression: Reliable prediction of the number of likes for social media posts.

Key Highlights

- Effective handling of class imbalance to achieve balanced classification performance.
- Modeling large fluctuations in regression targets with robust techniques like logarithmic transformations.
- Proper use of hyperparameter optimization to enhance model performance.

Recommendations

- Data Diversity: Increase dataset diversity and enrichment to further improve model generalizability.
- Alternative Approaches: Explore alternative modeling techniques, such as deep learning, for potential enhancements.