**Regression Pipeline for Predicting Post Engagement**

**Introduction**

This report outlines the development and execution of a comprehensive regression pipeline aimed at predicting the engagement level of social media posts, specifically the like\_count. The pipeline integrates advanced machine learning techniques, hyperparameter tuning, and feature engineering to optimize predictions. The workflow includes multiple stages, from loading and preprocessing data to training and evaluating various machine learning models. Additionally, a dedicated script facilitates predictions on unseen test data using a trained model.

**Data Loading and Preprocessing**

**Objective**

The data preprocessing stage prepares the dataset for modeling by transforming raw JSON data into a structured format suitable for machine learning algorithms.

**Steps**

1. **Loading Data:**
   * The dataset is extracted from a JSON file containing both user-level and post-level information.
   * Relevant features, such as follower\_count, following\_count, media\_type, and like\_count, are extracted from the raw data.
2. **Feature Engineering:**
   * A new feature, follower\_following\_ratio, is computed as the ratio of followers to followers plus one.
   * The media\_type categorical feature is transformed using one-hot encoding, resulting in columns such as media\_type\_0, media\_type\_1, and media\_type\_2.
   * Outliers in like\_count are capped at the 99th percentile to ensure robust model performance.
3. **Output:**
   * A cleaned and feature-engineered DataFrame, ready for model training, is produced.

**Model Training and Evaluation**

**Objective**

This stage involves training and tuning multiple machine learning models to identify the best-performing regressor for predicting like\_count.

**Steps**

1. **Model Selection:**
   * Four machine learning algorithms are evaluated: Random Forest, XGBoost, LightGBM, and CatBoost.
2. **Hyperparameter Tuning:**
   * Each model undergoes hyperparameter optimization using RandomizedSearchCV with 5-fold cross-validation.
   * Extensive parameter grids are defined for each model, allowing for fine-tuning of parameters such as n\_estimators, max\_depth, and learning rate.
3. **Evaluation Metrics:**
   * Models are evaluated on the test set using Mean Squared Error (MSE) and R² Score.
   * Cross-validation ensures robustness, and the best model is selected based on test performance.
4. **Pipeline Integration:**
   * Preprocessing steps, such as scaling numeric features, are integrated into a unified pipeline for consistency and efficiency.
5. **Results:**
   * The best-performing model is identified, and its hyperparameters are saved for future use.

**Prediction Workflow**

**Objective**

To predict the like\_count of posts in unseen test data using the trained model.

**Steps**

1. **Loading the Model:**
   * The trained model is deserialized from a pickle file (best\_model.pkl).
2. **Preparing Test Data:**
   * Test data is processed to align with the training dataset's structure.
   * Features such as time\_index and media\_type are computed or encoded.
   * Missing values are handled by defaulting to zero for numerical features.
3. **Prediction:**
   * The model's predict method generates predictions for the test dataset.
   * Results are stored as a dictionary mapping post\_id to predicted like\_count.
4. **Output:**
   * Predictions are saved as a JSON file (prediction-regression-round3.json) for downstream use.

**Key Features and Innovations**

1. **Feature Engineering:**
   * Introduction of the follower\_following\_ratio enhances the model's ability to capture user-level dynamics.
2. **Multi-Model Comparison:**
   * Evaluating four advanced models ensures the selection of the best-performing algorithm for the task.
3. **Comprehensive Pipeline:**
   * Integration of preprocessing, training, tuning, and evaluation within a single pipeline streamlines the workflow.
4. **Robust Test Data Preparation:**
   * Consistent processing of unseen data ensures reliable predictions.

**Challenges and Solutions**

**Handling Outliers**

Outliers in like\_count posed challenges during model training. Capping the like\_count at the 99th percentile mitigated their impact, ensuring stable model performance.

**Hyperparameter Optimization**

The large search space for hyperparameters increased computational costs. This was addressed by limiting the number of iterations in RandomizedSearchCV while maintaining comprehensive parameter grids.

**Conclusion**

This regression pipeline successfully predicts the like\_count for social media posts by leveraging advanced machine learning models and rigorous preprocessing techniques. The workflow ensures high accuracy and robustness through feature engineering, multi-model evaluation, and hyperparameter tuning. The pipeline's modular design allows for easy adaptation to similar prediction tasks in other domains.

**Future Work**

1. Incorporate additional features, such as textual analysis of captions, to further improve predictions.
2. Experiment with ensemble methods to combine the strengths of multiple models.
3. Optimize runtime performance by parallelizing certain preprocessing and training steps.

This pipeline serves as a strong foundation for future improvements and demonstrates the power of machine learning in predictive analytics.