**Project-4 Documentation**

**Feature Forge Enhancing Evaluating Ml Models**

**1. Overview**

The project aims to predict loan approvals based on customer information using machine learning models. It involves data exploration, feature engineering, model development, hyperparameter tuning, and deployment.

**2. Data Exploration and Cleaning**

* **Objective:** Understand the dataset structure and clean missing values.
* **Steps:**
  + Explored categorical and numerical features.
  + Visualized relationships between features and Loan\_Status.
  + Encoded categorical variables (e.g., mapping Loan\_Status to binary values Loan\_Status\_Y).

**3. Feature Engineering**

* **Objective:** Enhance dataset by creating new meaningful features.
* **New Features:**
  + Total\_Income = ApplicantIncome + CoapplicantIncome
  + EMI = LoanAmount / Loan\_Amount\_Term
  + Income\_to\_Loan\_Ratio = Total\_Income / LoanAmount

**4. Feature Selection**

* **Objective:** Identify the most important features for prediction.
* **Methods:**
  + **SelectKBest (ANOVA F-Statistic):**
    - Ranked features based on their statistical significance.
  + **RandomForest Feature Importance:**
    - Evaluated feature contributions using a Random Forest model.

**5. Dimensionality Reduction**

* **Objective:** Reduce feature dimensions while retaining variance.
* **Method:**
  + Principal Component Analysis (PCA) applied, reducing features to five principal components.

**6. Model Development**

* **Objective:** Train various machine learning models and evaluate their performance.
* **Models Implemented:**
  + Logistic Regression (with regularization)
  + Random Forest (with hyperparameter tuning via GridSearchCV)
  + XGBoost Classifier
  + Neural Network (MLPClassifier)
* **Cross-Validation Results:**
  + Logistic Regression achieved a mean accuracy of ~80%.
  + Random Forest's best cross-validation score was ~81%.
  + XGBoost and Neural Network achieved ~73% and ~72%, respectively.

**7. Hyperparameter Tuning**

* **Objective:** Optimize model performance by fine-tuning parameters.
* **Methods:**
  + GridSearchCV applied to Random Forest for parameters like n\_estimators, max\_depth, min\_samples\_split, and min\_samples\_leaf.

**8. Stress Testing**

* **Objective:** Validate model robustness under noisy data conditions.
* **Results:**
  + Logistic Regression maintained an accuracy of ~80% when tested with noisy data.

**9. Performance Evaluation**

* **Objective:** Evaluate the final model's performance on test data.
* **Metrics:**
  + **Accuracy:** 86.18%
  + **Precision:** 84.00%
  + **Recall:** 98.82%
  + **F1 Score:** 90.81%
* **Classification Report:**
  + High recall indicates the model effectively predicts loan approvals (Loan\_Status = 1).

**10. Deployment Pipeline**

* **Objective:** Prepare the model for real-world deployment and monitoring.
* **Steps:**
  + **Model Saving:** Saved using joblib with versioning.
  + **Monitoring:** Simulated CI/CD pipeline to evaluate predictions on new data batches.
  + **Logging:** Model performance logs stored for monitoring and debugging.