Feature Forge: Enhancing & Evaluating ML Models

Dataset: Loan Prediction Dataset (Analytics Vidhya)

Goal: To develop a predictive machine learning model for loan approval decisions based on applicant data.

Milestone 1: Feature Engineering & Selection

An extensive exploratory data analysis (EDA) was conducted at the start of the project in order to comprehend the distribution and structure of the dataset. To show skewness and variance, histograms were made for numerical variables like loan amount, loan term, coapplicant income, and applicant income. To investigate class distributions and their relationship to loan status, count plots were used to visualize categorical variables such as gender, marital status, number of dependents, education, self-employment status, credit history, and property area. A correlation heatmap indicated characteristics that were more closely linked to the loan approval outcome and assisted in determining relationships between numerical variables.

By merging applicant and coapplicant income into a single metric that more accurately reflects the household's overall earning power, a new feature named TotalIncome was added to the dataset using feature engineering. In order to prepare the data for model training, the preprocessing pipeline also included handling missing values appropriately and using encoding techniques to convert categorical features into numerical format.

Milestone 2: Model Enhancement & Tuning

To create a precise predictive model, several classification algorithms were investigated. The processed dataset was used to train and assess the Gradient Boosting Classifier, Random Forest Classifier, and Logistic Regression. These models were trained on the feature-engineered dataset with either default or tuned hyperparameters, and were selected based on their efficacy in handling classification tasks. Every model contributed a different method for extracting patterns from the data, which served as a foundation for comparison in order to determine the best strategy.

Milestone 3: Model Evaluation

Industry-standard evaluation metrics were used to assess the trained models' performance. Accuracy, precision, recall, and F1-score for each of the predicted classes were evaluated in a

classification report. To visually examine the balance of true positives, false positives, true negatives, and false negatives, a confusion matrix was employed. To assess model discrimination, which is crucial for unbalanced datasets like loan approval predictions, ROC AUC and F1 scores were also computed. These metrics gave a thorough picture of each model's performance in forecasting loan results.

Conclusion

This project effectively illustrated how to solve a real-world classification problem using data exploration, feature engineering, and machine learning modeling techniques. A strong basis for creating an automated loan prediction system was established by carefully preparing and improving the dataset and experimenting with various classification models. The entire process demonstrates a methodical and efficient approach to practical machine learning tasks.