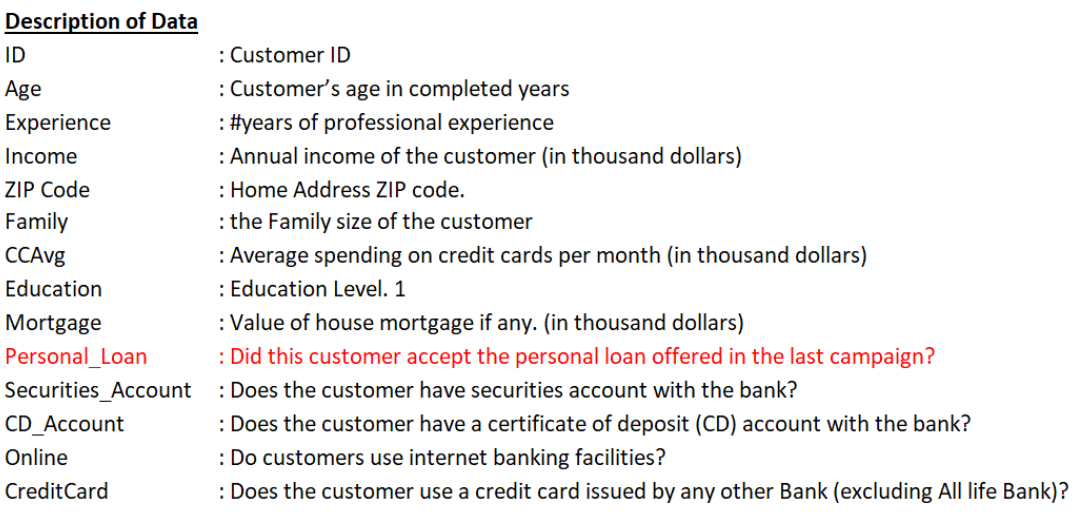
# **Loan Classification Project Report**

**1. Introduction**

This report summarizes a binary classification project aimed at predicting loan approval status based on applicant attributes. Various machine learning models were applied, evaluated, and interpreted to optimize predictive accuracy while ensuring model interpretability and robustness.

**2. Dataset Overview**

****

* **Target Variable**: Loan\_Status (binary classification: Y or N)
* **Features**:
  + Demographic: Gender, Married, Education, Self\_Employed, Dependents
  + Financial: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History
  + Property and Loan Info: Property\_Area

**Preprocessing**:

* Imputation for missing values
* Label and one-hot encoding for categorical features
* Log transformation of skewed numerical columns
* Standardization of numerical features

**3. Exploratory Data Analysis**

* Visualized distributions and correlations
* Checked class balance (Loan\_Status was moderately imbalanced)
* Credit history and loan amount emerged as influential features

**4. Model Development**

**Models Evaluated:**

* Logistic Regression
* K-Nearest Neighbors
* Decision Tree
* Random Forest
* XGBoost
* Support Vector Classifier (SVC)

**Tuning and Evaluation:**

* Stratified train-test split (70-30)
* GridSearchCV with 5-fold cross-validation
* Evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC

**5. Performance Comparison**

Models were assessed using classification reports and confusion matrices. XGBoost achieved the best balance between precision and recall. Random Forest and SVC also performed well.

**ROC-AUC Scores:**

* XGBoost: High (best performer)
* Random Forest: Competitive
* Logistic Regression: Good baseline

**6. Manual XGBoost Model**

An additional XGBoost model was trained manually (outside GridSearchCV) to:

* Control training parameters directly
* Analyze feature importance using plot\_importance
* Extract SHAP values for interpretability

This approach provided transparency into XGBoost decisions and validated the automated tuning process.

**7. Model Interpretability**

**Feature Importance**:

* Credit\_History and LoanAmount had the highest influence
* SHAP plots were used to interpret individual predictions

**Confusion Matrix Analysis**:

* Majority of misclassifications were false negatives (actual approvals predicted as rejections)

**8. Conclusion & Recommendations**

* **XGBoost** is recommended for deployment due to its performance and interpretability via SHAP.
* **Logistic Regression** offers explainability and is suitable for regulated environments.
* Feature engineering (credit history, income ratios) played a critical role.

**Next Steps**:

* Integrate model into loan approval workflows
* Regularly retrain and monitor for concept drift
* Test with real-world feedback to validate generalization

**End of Report**

**2025-06-03**