Final Report: Exploratory Data Analysis of the Home Loan Dataset

Project: Home Loan Dataset EDA

Date: 03/11/25

Author: Dasaolu Samuel

1. Project Overview and Objective

The objective of this project was to conduct a comprehensive Exploratory Data Analysis (EDA) on the Home Loan dataset. The primary goal was to understand the underlying structure of the data, identify patterns, and find key relationships that influence loan approval status. This analysis is essential for uncovering insights that can support subsequent predictive modeling and strategic decision-making in home loan processing.

2. Methodology (The EDA Process)

Phase 1: Data Collection and Preparation

- **Loaded Data:** The `home_loan_train.csv` and `home_loan_test.csv` datasets were loaded into pandas DataFrames.
- **Inspected Data:** The training data was inspected, revealing missing values in `Gender`, `Married`, `Dependents`, `Self_Employed`, `LoanAmount`, `Loan_Amount_Term`, and `Credit_History`. It also revealed a non-numeric `Dependents` column (due to '3+').
- **Data Cleaning:**
  - **Categorical Imputation:** Missing values for all categorical columns (including `Credit_History`) were filled using the **mode** (most frequent value).
  - **Numerical Imputation:** Missing `LoanAmount` was filled with the **median** (to resist outlier skew) and `Loan_Amount_Term` was filled with the **mode**.
  - **Inconsistency Correction:** The `Dependents` column was cleaned by replacing '3+' with '3' and converting the column to a numeric data type.

Phase 2: Exploratory Data Analysis

- **Descriptive Analysis:** We analyzed the distributions of all numerical and categorical features.
    - **Numerical:** `ApplicantIncome`, `CoapplicantIncome`, and `LoanAmount` were all found to be **heavily right-skewed** with significant outliers.
    - **Categorical:** `Credit_History` (1.0) was the dominant class (85%), as were `Male` (81%), `Married` (65%), and `Graduate` (78%). `Property_Area` was well-distributed, with `Semiurban` being the most common.
- **Relationship Analysis:** We examined the relationship between each feature and the target variable, `Loan_Status`, using three methods:
    1. **Cross-tabulations:** Used for categorical features, most notably showing the powerful link between `Credit_History` and `Loan_Status`.
    2. **Scatter Plots:** Used for numerical features (e.g., `Total_Income` vs. `LoanAmount`), which showed no clear linear separation between approved and rejected loans.
    3. **Correlation Matrix:** This numerically confirmed the strong positive correlation (**0.54**) between `Credit_History` and `Loan_Status` and the weak correlation for all other numerical features.

---

3. Key Findings and Insights

- **Finding 1: Credit History is the Most Dominant Predictor.**

  This is the most significant finding. Our cross-tabulation revealed that applicants with a good credit history (1.0) were **approved 79%** of the time. Conversely, applicants with a bad credit history (0.0) were **rejected 92%** of the time. This suggests a good credit history is a near-mandatory prerequisite for loan approval.

- **Finding 2: Property Area is a Strong Secondary Predictor.**

  The location of the property is a key factor. Applicants from `Semiurban` areas have the highest approval rate, followed by `Urban` areas. `Rural` areas have the lowest approval rate.

- **Finding 3: Marital Status Shows a Clear Trend.**

  Married applicants have a noticeably higher approval rate than non-married applicants.

- **Finding 4: Income/Loan Amount is Not a Simple Differentiator.**

  A key finding was the *lack* of a simple, direct relationship between income (or loan amount) and approval. Our scatter plots showed that approved and rejected loans exist at all income levels. This implies that while income is part of the equation, it is not a simple "high income = approval" rule and is likely evaluated *in combination* with other factors (like `Credit_History`).

---

4. Recommendations for Next Steps (Predictive Modeling)

1. **Feature Engineering:**
   - **Handle Skewness:** The income and loan amount features are extremely skewed. A **log transform** should be applied to `ApplicantIncome`, `CoapplicantIncome`, and `LoanAmount` to normalize their distributions, which will benefit most models.
   - **Create New Features:** Consider creating a `Total_Income` feature (`ApplicantIncome` + `CoapplicantIncome`) and a `Debt_to_Income_Ratio` feature (`LoanAmount` / `Total_Income`) as these may be more predictive.
2. **Modeling Strategy:**
   - **Handle Imbalance:** The target variable `Loan_Status` is imbalanced (69% 'Y' vs. 31% 'N'). Techniques such as **SMOTE** (oversampling the 'N' class) or using a model with a `class_weight='balanced'` parameter will be necessary to prevent the model from just predicting 'Y'.
   - **Prioritize Features:** `Credit_History` and `Property_Area` should be considered high-importance features in any model.