Exploratory Data Analysis Report

Insights from Loan Approval Data

Prepared by: RetentionX

May 2025

Abstract

This report presents an exploratory data analysis (EDA) of a loan approval dataset, uncovering critical patterns and relationships among demographic, financial, and behavioral features. Key predictors of loan approval, such as income and credit score, are identified, and data quality issues are addressed to prepare for predictive modeling. The findings offer actionable insights for financial institutions to streamline loan approval processes.

Contents

1 Data Overview

The dataset comprises **5,000 rows** and **12 columns**, capturing demographic, financial, and loan-related information. Below is a summary of the features:

- **Numerical Features (5)**:
  - `age`: Age of the individual (integer)
  - `income`: Annual income in USD (float)
  - `credit_score`: Credit score (integer)
  - `loan_amount`: Loan amount applied for (float)
  - `account_balance`: Current account balance (float)
- **Categorical Features (7)**:
  - `gender`: Gender (Male/Female)
  - `education`: Education level
  - `marital_status`: Marital status
  - `occupation`: Occupation type
  - `city`: City of residence
  - `loan_status`: Loan approval status (Approved/Rejected)
  - `dependents`: Number of dependents (integer, treated as categorical)

2 Data Quality Assessment

2.1 Missing Values

- `income`: 2.4% missing (120 records)
- `credit_score`: 1.1% missing (55 records)
- Other features: No missing values
- **Treatment**: Imputed missing values with median to preserve distribution

2.2 Duplicates

- 8 duplicate records identified
- **Handling**: Removed duplicates to ensure data integrity

3 Statistical Summary

3.1 Numerical Features

Feature	Mean	Median	Std	Q1	Q3
age	37.2	36	10.5	29	45
income	$48,500	$45,000	$15,200	$35,000	$58,000
credit_score	690	700	50	660	730
loan_amount	$12,300	$10,000	$7,800	$6,000	$15,000
account_balance	$5,200	$4,800	$2,100	$3,500	$6,500

Table 1: Statistical summary of numerical features

3.2 Categorical Features

Feature	Distribution
gender	52% Male, 48% Female
education	40% Graduate, 35% Postgraduate, 25% High School
marital_status	60% Married, 40% Single
loan_status	68% Approved, 32% Rejected

Table 2: Distribution of categorical features

4 Data Distribution Analysis

- age, income, and loan_amount: Right-skewed distributions
- Outliers:
  - income: Top 1% (values > $80,000)
  - loan_amount: Top 2% (values > $25,000)
- credit_score: Approximately normal (skewness = -0.1, kurtosis = 2.8)
- account_balance: Mild positive skewness

5 Feature Relationships

- **Correlation Analysis**:
  - income and loan_amount: $r = 0.62$ (moderate positive)

  - `credit_score` and `loan_status`: $r = 0.48$ (positive)
  - `age` and `account_balance`: $r = 0.30$ (weak positive)
- **Key Patterns**:
  - Higher `income` and `credit_score` strongly predict loan approval
  - `education` level correlates with higher `income`
  - Married applicants have slightly higher approval rates

6 Preprocessing Decisions

6.1 Data Cleaning

- Imputed missing `income` and `credit_score` with median
- Removed 8 duplicate records
- Capped outliers in `income` and `loan_amount` at 99th percentile
- Applied Min-Max scaling to `income`, `loan_amount`, and `account_balance`

6.2 Feature Engineering

- Created `income_per_dependent` = `income` / (`dependents` + 1)
- One-hot encoded categorical variables: `education`, `occupation`, `city`
- Selected top 8 features based on feature importance analysis

7 Conclusions

7.1 Key Insights

- `income` and `credit_score` are the strongest predictors of `loan_status`
- Data quality issues (missing values, outliers) were minimal and effectively handled
- Feature engineering enhanced model readiness

7.2 Recommendations

- Consider binning `age` and `income` for improved model robustness

- Monitor categorical feature distributions for data drift
- Address `loan_status` class imbalance using resampling techniques

7.3 Potential Challenges

- Imbalanced `loan_status` (68% Approved vs. 32% Rejected) may require SMOTE or oversampling
- High cardinality in `city` could complicate modeling