Project Title: Consumer Lending Risk Insights Through Data-Driven Analytics

Author: Mehraan Khan

PRN: 250840325042 **Date:** 07/12/2025

---

1. Executive Summary

This case study aims to identify key drivers of consumer loan defaults using a dataset of 307,511 loan applicants. Through extensive data cleaning, exploratory data analysis (EDA), and statistical hypothesis testing, we assessed the relationship between applicant demographics, financial profiles, and loan repayment behavior.

Key Findings:

- **Age is a critical factor:** Defaulters are significantly younger (Average ~40.7 years) compared to non-defaulters (~44.3 years).
- **Financial Strain:** Applicants who default tend to have lower total incomes but higher annuity (EMI) obligations.
- **History Matters:** There is a statistically significant relationship between having a previous loan application refused and defaulting on a current loan.
- **Demographics:** Gender and Education level show significant variances in default rates.

---

2. Introduction

2.1 Objective

The primary objective of this analysis is to minimize financial risk for the lending institution by identifying patterns and characteristics associated with "Target 1" (clients with payment difficulties).

2.2 Data Description

The analysis utilizes two datasets:

1. **Application Data:** Current loan application details (Income, Credit Amount, Family Status, etc.).
2. **Previous Application Data:** History of previous loans and application statuses.

Initial Dataset Size: 307,511 rows, 122 columns.

---

3. Data Preprocessing & Methodology

To ensure data quality and analytical accuracy, the following preprocessing steps were undertaken:

3.1 Data Cleaning

- **Feature Selection:** Reduced high-dimensionality data to key columns involving demographics (Age, Gender), financials (Income, Credit, Annuity), and external scores.
- **Missing Values:**
    - Numerical columns (e.g., `AMT_INCOME_TOTAL`, `AMT_CREDIT`) were imputed using the **Median** strategy to mitigate skewness.
    - Categorical columns were imputed using the **Mode**.
    - Anomalies in `DAYS_EMPLOYED` (e.g., 365243 days) were treated as null values and imputed.

3.2 Outlier Treatment

- The Interquartile Range (IQR) method was applied to remove extreme outliers in Income, Credit, and Annuity columns to prevent skewed statistical results.

3.3 Feature Engineering

- **Age Conversion:** Converted `DAYS_BIRTH` to `AGE_YEARS`.
- **Employment Duration:** Converted `DAYS_EMPLOYED` to `EMPLOYED_YEARS`.
- **Previous Refusal Flag:** Merged with the previous application dataset to create a binary feature `HAS_PREV_REFUSAL`, indicating if the client had been rejected for a loan in the past.

---

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

Distributions of key variables revealed the following:

- **Income & Credit:** Both distributions were right-skewed, indicating a high volume of lower-to-middle income applicants.
- **Age:** The age distribution is fairly normal, but younger applicants appear more frequently in the default category.

4.2 Bivariate Analysis

- **Income vs. Default:** A Heatmap analysis of Income Bins vs. Credit Bins revealed that lower income levels combined with high credit amounts pose higher risks.

- **Education:** Applicants with "Secondary/Secondary Special" education showed higher default counts compared to those with "Higher Education."
- **Occupation:** Laborers and Sales staff represent the highest volume of applicants and also the highest absolute number of defaults.

---

5. Statistical Hypothesis Testing

To validate visual observations, statistical tests (T-Tests for numerical data and Chi-Square for categorical data) were conducted at a 95% confidence interval (p-value threshold: 0.05).

Test 1: Income vs. Default

- **Hypothesis (H0):** There is no difference in mean income between defaulters and non-defaulters.
- **Test Used:** T-Test
- **Result:** P-value < 0.05
- **Conclusion: Reject Null Hypothesis.** Defaulters have a significantly lower mean income (~149k) compared to non-defaulters (~151k).

Test 2: Gender vs. Default

- **Hypothesis (H0):** Default is independent of gender.
- **Test Used:** Chi-Square Test
- **Result:** P-value ≈ 0.0
- **Conclusion: Reject Null Hypothesis.** There is a statistically significant dependency between gender and default rates.

Test 3: Age vs. Default

- **Hypothesis (H0):** Mean age of defaulters is equal to non-defaulters.
- **Test Used:** T-Test
- **Result:** P-value = 0.0
- **Conclusion: Reject Null Hypothesis.** Defaulters are significantly younger (Avg 40.7 years) than non-defaulters (Avg 44.3 years).

Test 4: Previous Refusals vs. Default

- **Hypothesis (H0):** Previous loan refusals are independent of current default.
- **Test Used:** Chi-Square Test
- **Result:** P-value < 0.05
- **Conclusion: Reject Null Hypothesis.** Clients who have been refused a loan in the past are statistically more likely to default on current loans.

Test 5: Annuity Amount vs. Default

- **Hypothesis (H0):** Mean annuity is equal for both groups.

- **Test Used:** T-Test
- **Result:** P-value < 0.05
- **Conclusion: Reject Null Hypothesis.** Defaulters tend to have higher annuity payments (Avg 25.2k) compared to compliant payers (24.6k), suggesting higher financial burden.

---

6. Conclusion & Recommendations

6.1 Risk Profile Summary

The analysis confirms that the "Default" target variable is not random but correlated with specific demographic and financial traits. The highest risk profile includes:

1. **Younger Applicants:** Especially those under 40.
2. **Financial Stress:** Applicants with lower income but higher annuity obligations.
3. **Historical Behavior:** Applicants with a history of loan refusals.

6.2 Recommendations

1. **Age-Based Risk Adjustment:** Implement stricter underwriting criteria or required collateral for applicants under the age of 30-35.
2. **Debt-to-Income Scrutiny:** Focus on the Annuity-to-Income ratio rather than just total credit amount, as high annuities are a strong predictor of default.
3. **History Integration:** Ensure the credit scoring model heavily weights the `HAS_PREV_REFUSAL` flag, as past rejections are a strong signal of current risk.