Credit Risk Analysis: Exploring Loan Default Patterns in Banking Data



1. Introduction

Credit risk assessment is one of the most crucial components in the banking sector. Predicting whether a customer is likely to default on a loan helps financial institutions minimize losses and maintain financial stability. In this blog, we delve into a real-world bank dataset containing customer demographic and financial information to uncover patterns and relationships that impact loan default probability.

2. Dataset Overview

Our dataset contains **18 columns**, including both categorical and numerical variables:

Categorical Columns:

1. Family Status (e.g., Married, Single, Divorced)
2. Underage Children Count
3. Industry Name (e.g., Finance, Healthcare, Public Administration)
4. Credit History Rating (e.g., A1, B2, C2)
5. Loan Term (duration in months)
6. Education (e.g., Higher, Secondary, Secondary Special)
7. Sex (Male/Female)

Numerical Columns:

1. Debt (current credit debt amount)
2. Initial Limit (credit limit at origination)
3. Income (customer’s monthly income)
4. Probability of Default (PD) (ranging from 0 to 1)
5. Scoring Mark (credit score)

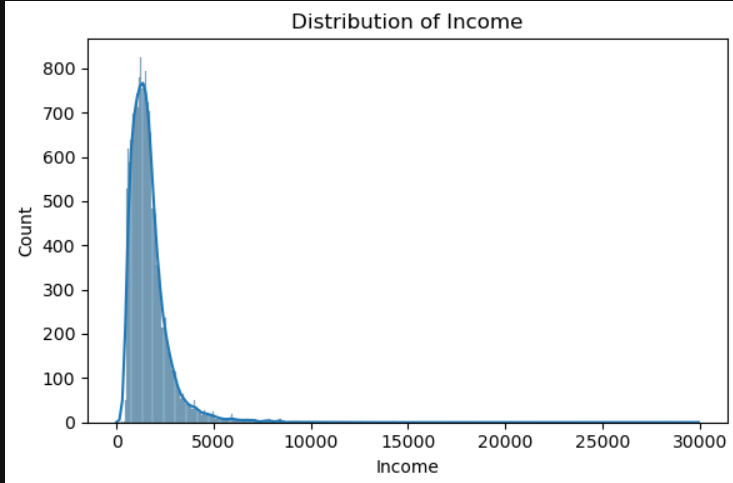
The primary focus is the Probability of Default (PD), a key metric that quantifies the likelihood of a customer defaulting on a loan.

3. Univariate Analysis

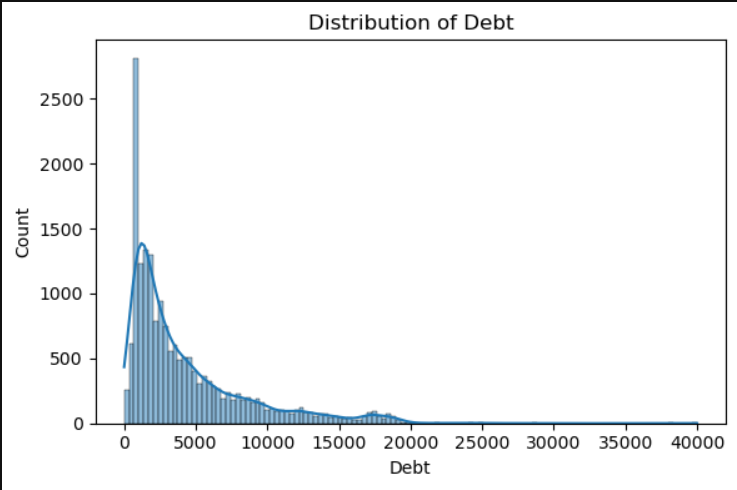
We began by examining individual column distributions to understand their spread and detect anomalies.

3.1. Key Observations:

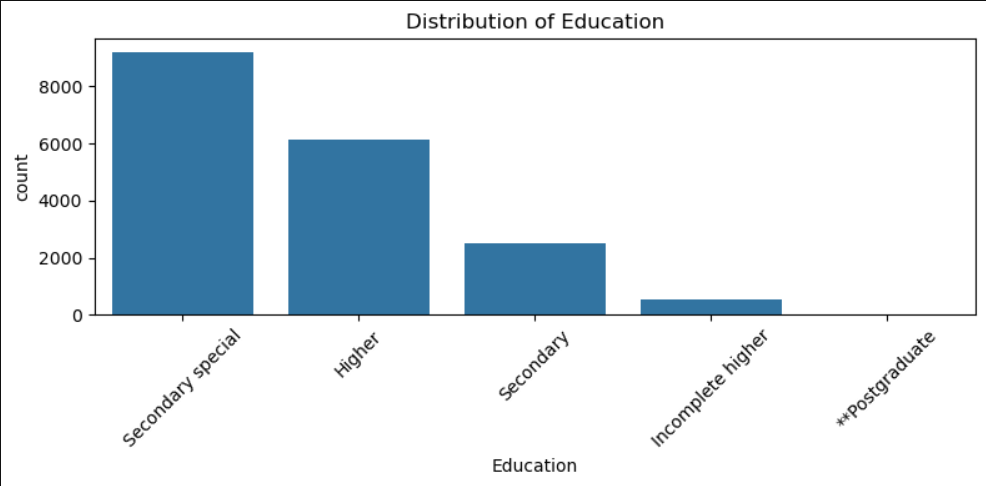
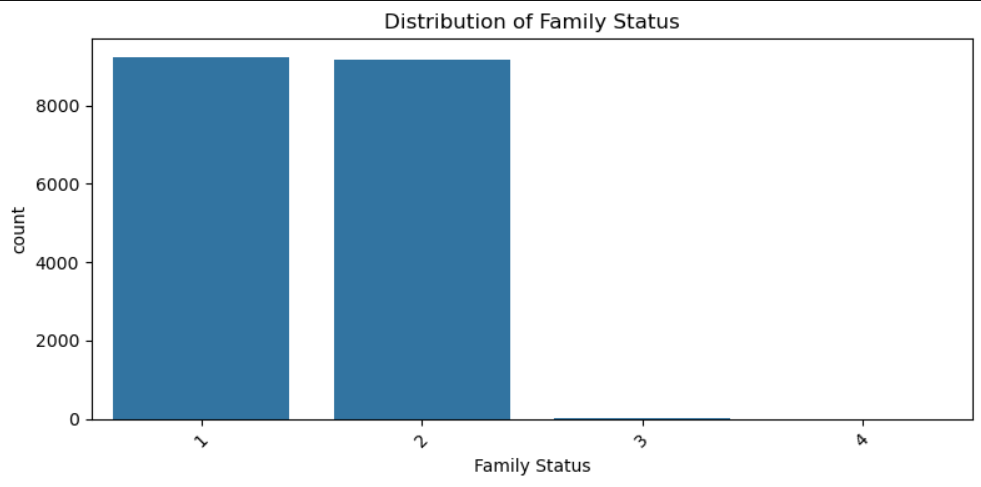
1. Income is right-skewed, indicating a few high earners.



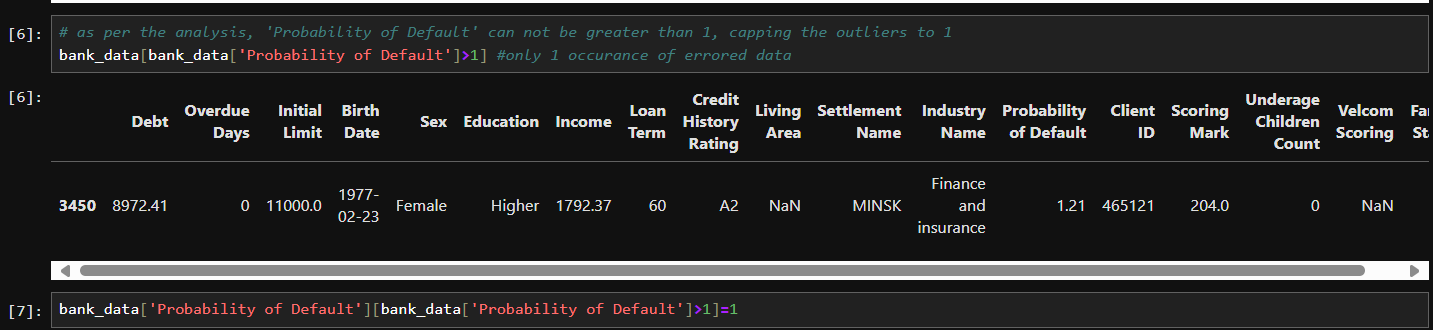
2. Debt varies significantly but clusters at lower ranges.



3. Education and Family Status show clear dominance in certain categories (e.g., most customers have "Secondary Special" education).

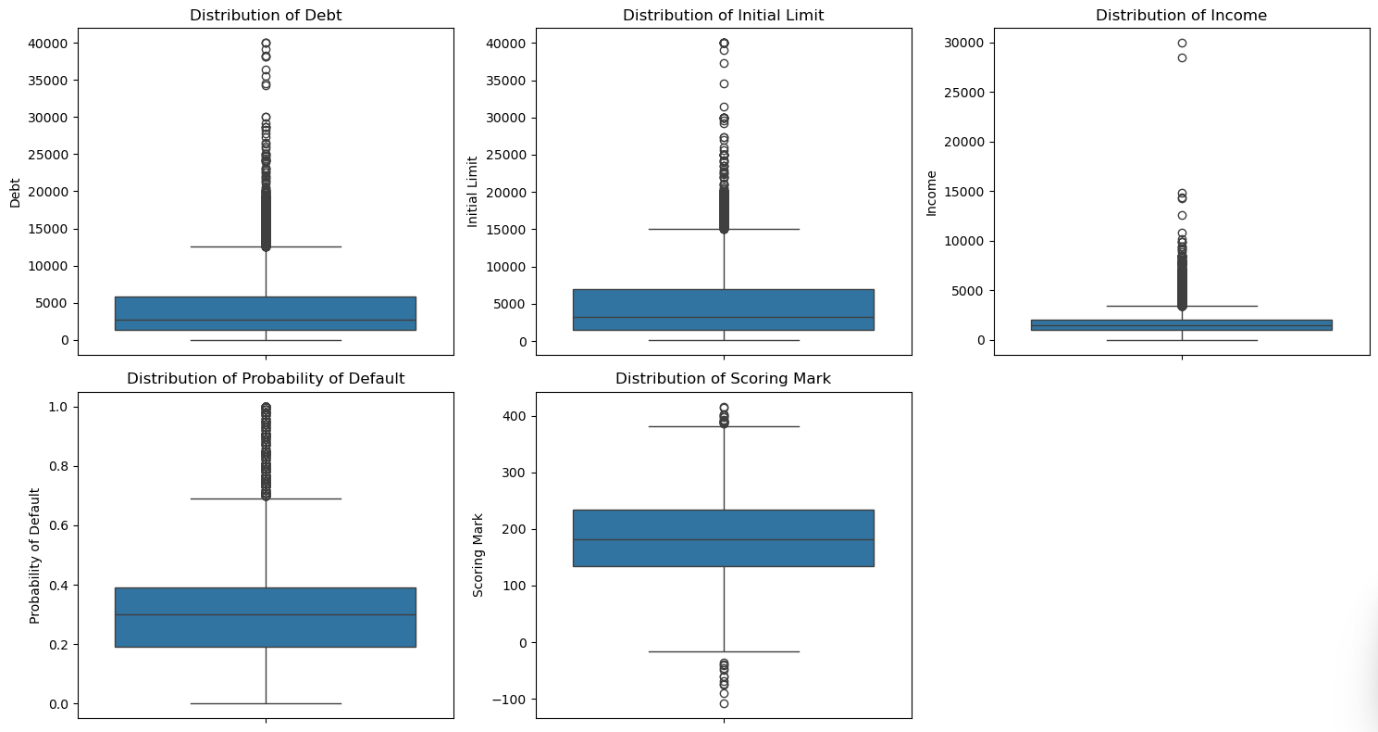
 

4. Probability of Default is capped at 1, ensuring realistic risk assessment.



3.2. Visualizations Used:

1. Boxplots for numerical columns (identifying outliers in Debt, Income).



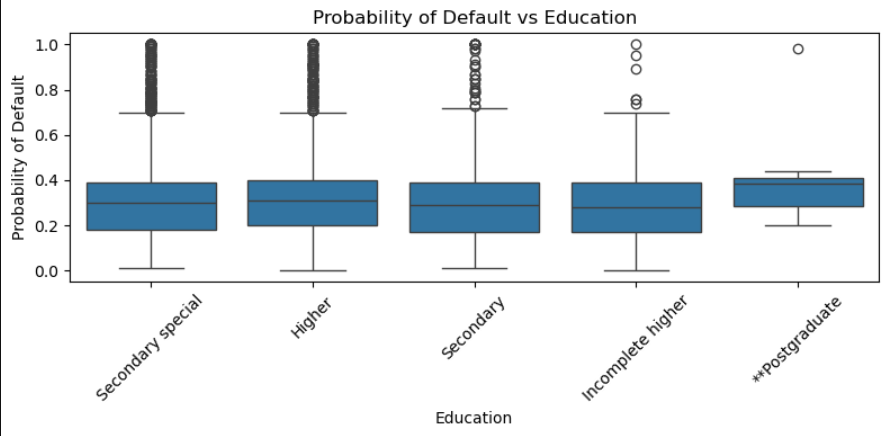
2. Bar plots for categorical distributions (e.g., Education levels, Family Status).

4. Bivariate Analysis

Next, explored relationships between features and Probability of Default.

1. Categorical vs. Probability of Default
   1. Education:

* Higher education correlates with lower default risk.



* Secondary education shows higher PD.
  1. Credit History Rating:
* A1 (best rating)→ Lowest PD.
* C2 (worst rating) → Highest PD.
  1. Loan Term:
* Longer loan terms (60 months) have higher PD than shorter terms (36 months).
  1. Sex:
* Minimal difference in PD between genders.

1. Numerical vs. Probability of Default
   1. Scatterplots & Correlation Heatmaps revealed:

* Weak correlation between Debt/Income and PD.
* Strong inverse relationship between Scoring Mark and PD (better scores = lower risk).

1. Multivariate Insights

We layered in complexity to uncover deeper trends:

1. Certain industries (e.g., Agroindustrial) have higher default risk.
2. Credit History Rating remains the strongest predictor, even when segmented by Family Status or Loan Term.
3. Customers with lower income and longer loan terms tend to have higher PD.

Visualizations Used:

1. Pairplots (joint distributions of key features).
2. Boxen plots (segmented PD by multiple categories).
3. Conclusion
4. Credit History Rating is the most influential predictor of default.
5. Education and Loan Term also play significant roles.
6. Demographic factors (Sex, Underage Children Count) show limited standalone predictive power.

Final Thoughts

Data science in banking isn’t just about predicting defaults, it’s about transparent, fair, and data-driven decisions that benefit both banks and customers. By refining our models with domain insights, we can enhance risk assessment and build trust in financial systems.

---------------------------------------------------------THANK YOU---------------------------------------------------