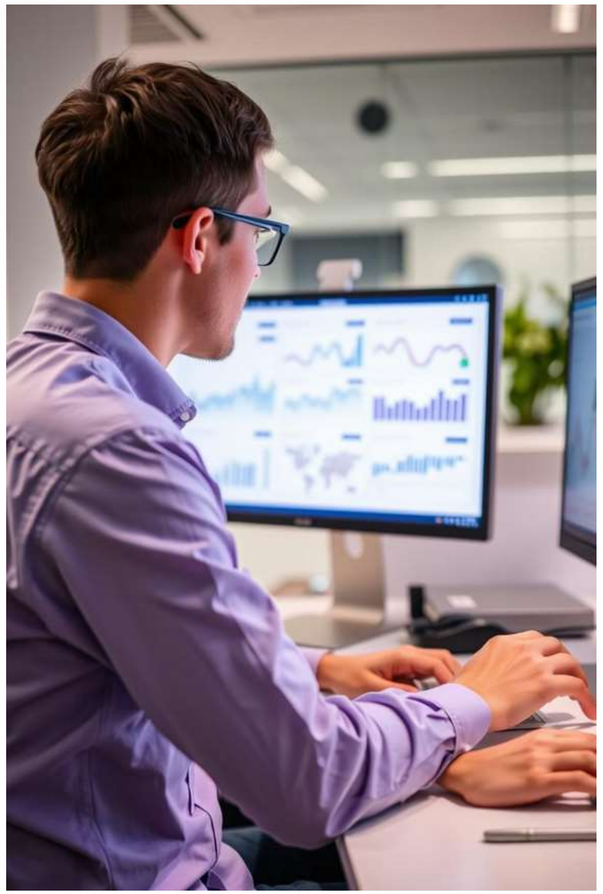
Exploratory Data Analysis (EDA) in banking

Discover how EDA transforms raw banking data into actionable insights for better loan decisions.

By;

Sohini

Vaibhavi B Raj



Introduction to EDA

What is EDA?

Exploratory Data Analysis helps understand dataset structure, patterns, and relationships before modeling.

Early Problem Detection

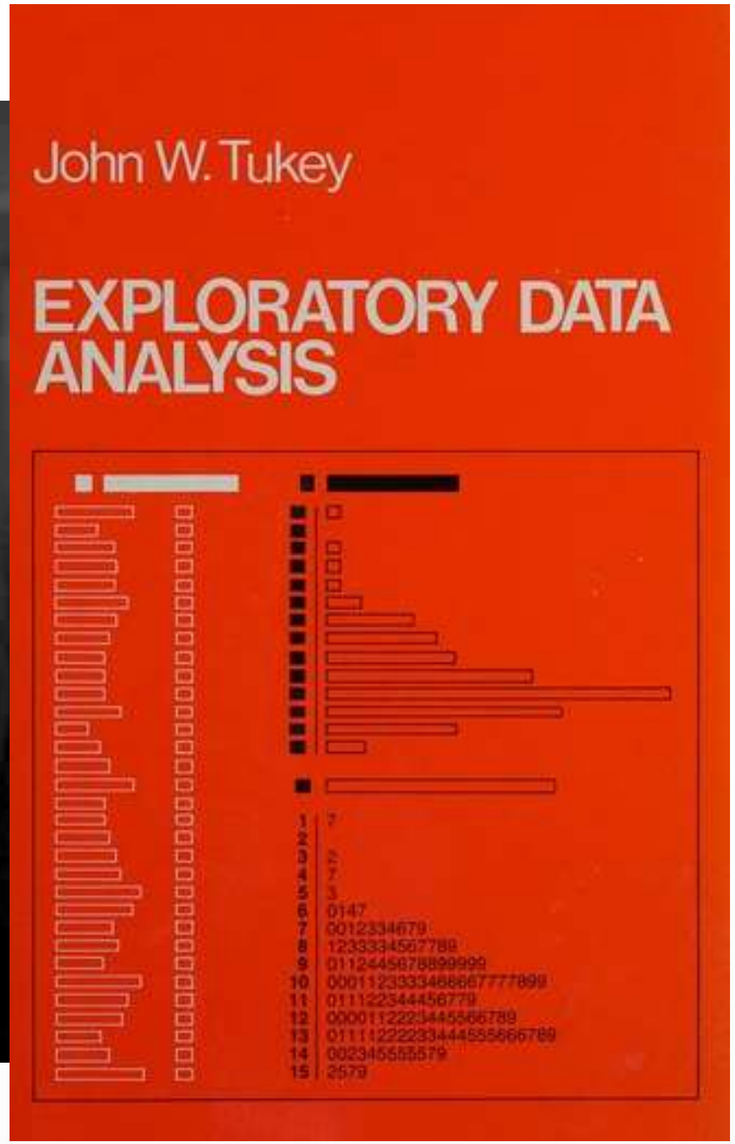
EDA identifies issues in data that could affect analysis quality.

- Maximise the insight in the data set
- Detect outliers and anomalies
- Test underlying assumptions

Better Decision-Making

It provides insights that improve business and modeling decisions.

History of EDA



01 Work of Tukey:1977

02 ABC’S of EDA, Velleman and Hoaglin(1981)

03 More graphical analysis has been developed lately

How EDA Helps in Jobs

Spot Insights

Extract valuable patterns from raw data.

Build Better Models

Create more accurate predictive models.

Improve Reporting

Enhance quality of data presentations.

Support Decisions

Provide data-backed business recommendations.

Essential skill for Data Analysts, Business Analysts, and Risk Analysts.

Where is EDA Applicable

Banking

Risk profiling and fraud detection

Insurance

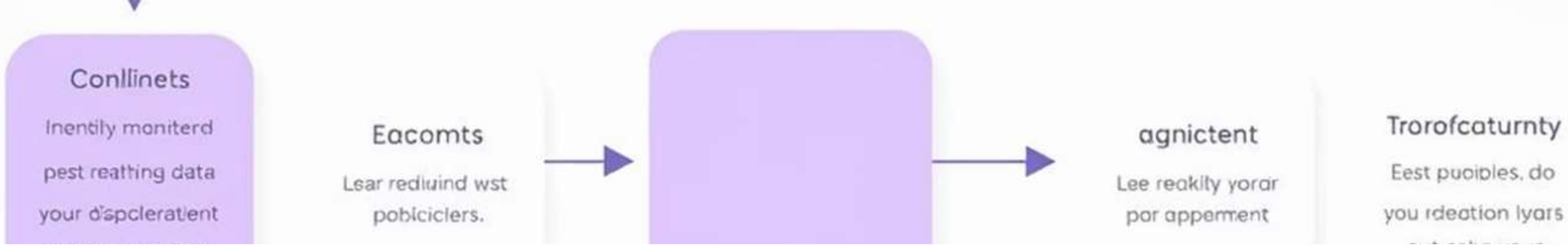
Claims assessment and risk underwriting

Healthcare

Patient history and risk prediction

Retail

Customer segmentation and demand forecasting



EDA Step-by-Step Visual

Business Understanding

Data Collection

Data Cleaning

Univariate Analysis

Bivariate Analysis

Outlier Detection

Correlation Study

Insights

Problem Statement

Risk Analytics

Assess credit risk in loan applications.

Customer Types

With and without payment difficulties.

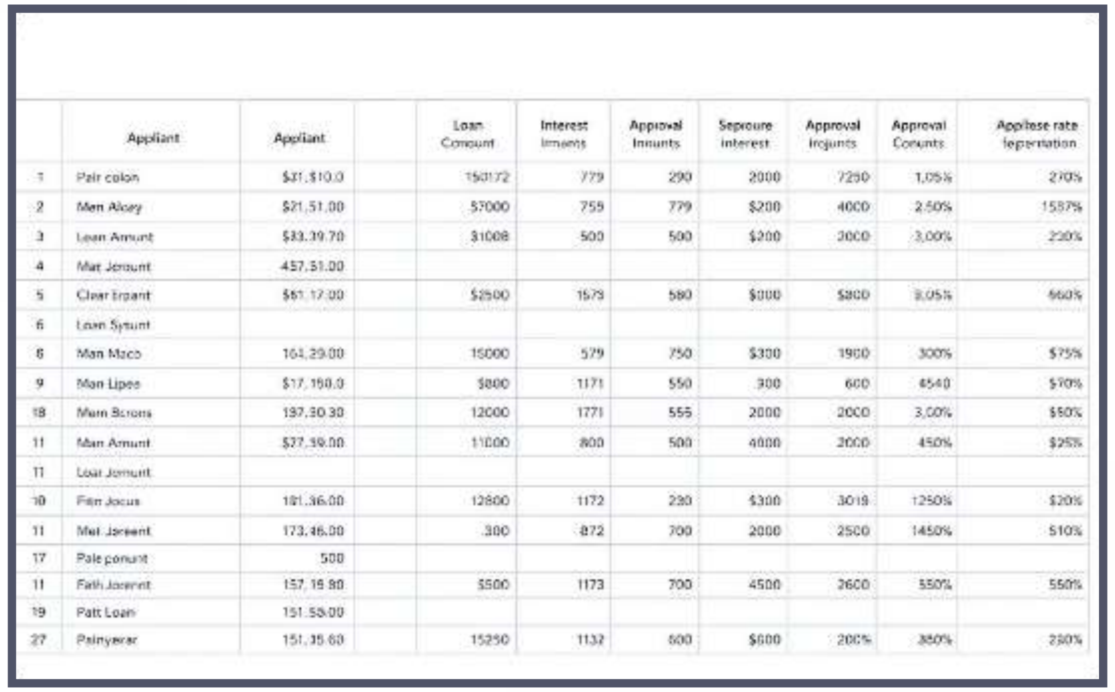
Decision Impact

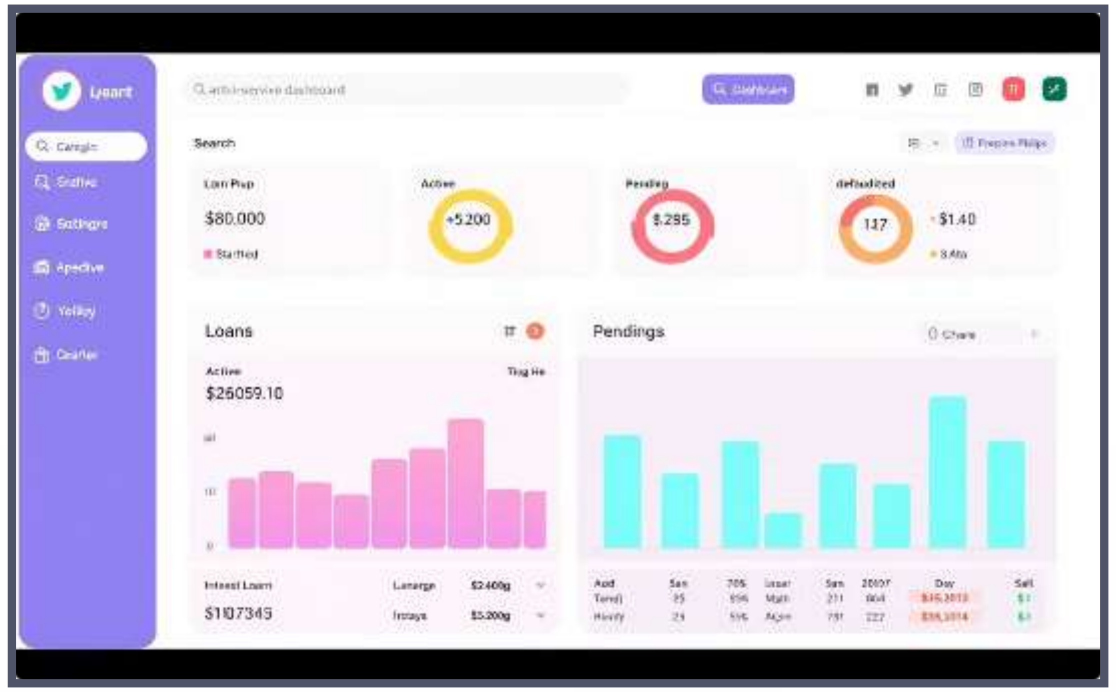
Rejecting good customers vs. approving risky ones.

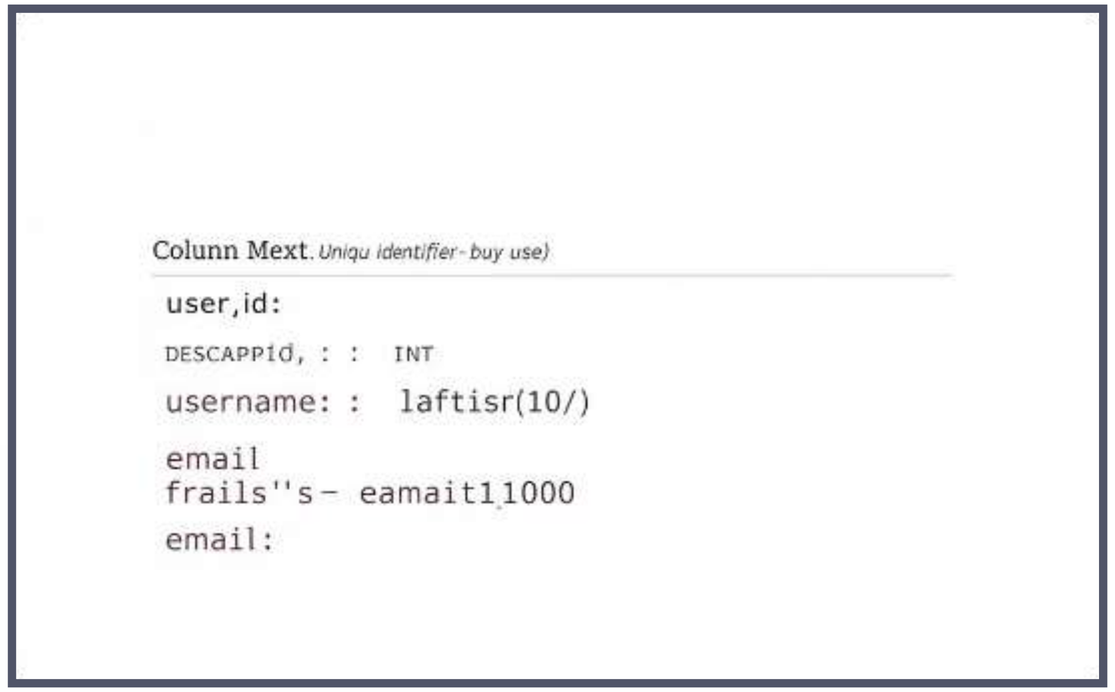
Loan Outcomes

Approved, Refused, Cancelled, Unused Offer.

Datasets







Application Data

Current application details with customer information and loan specifics.

Previous Applications

Customer's past loan application history and outcomes.

Data Dictionary

Detailed explanation of all variables in the datasets.

https://www.kaggle.com/datasets/sid9300/credit-eda-case-study-data

Approach and Methodology

Data Cleaning

Handled missing data, dropped high-NaN columns (>19%), fixed negatives.

Categorical Fixes

Replaced invalid entries like ‘XNA’ with logical values.

Univariate & Bivariate Analysis

Studied variable distributions and relationships.

Segmentation

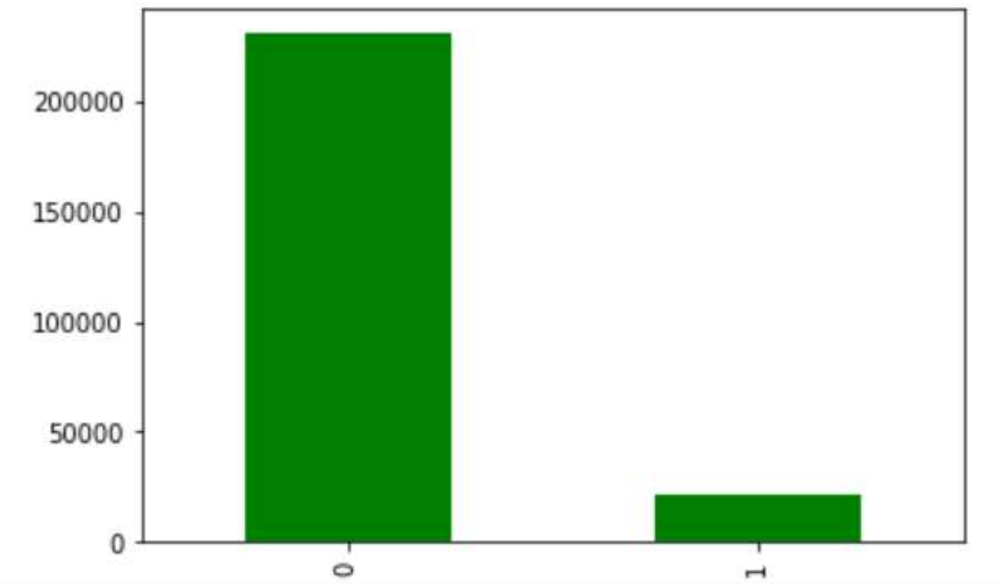
Separate analysis for defaulters vs non-defaulters.

Correlation & Outliers

Heatmaps to detect key factors; outliers observed via boxplots.

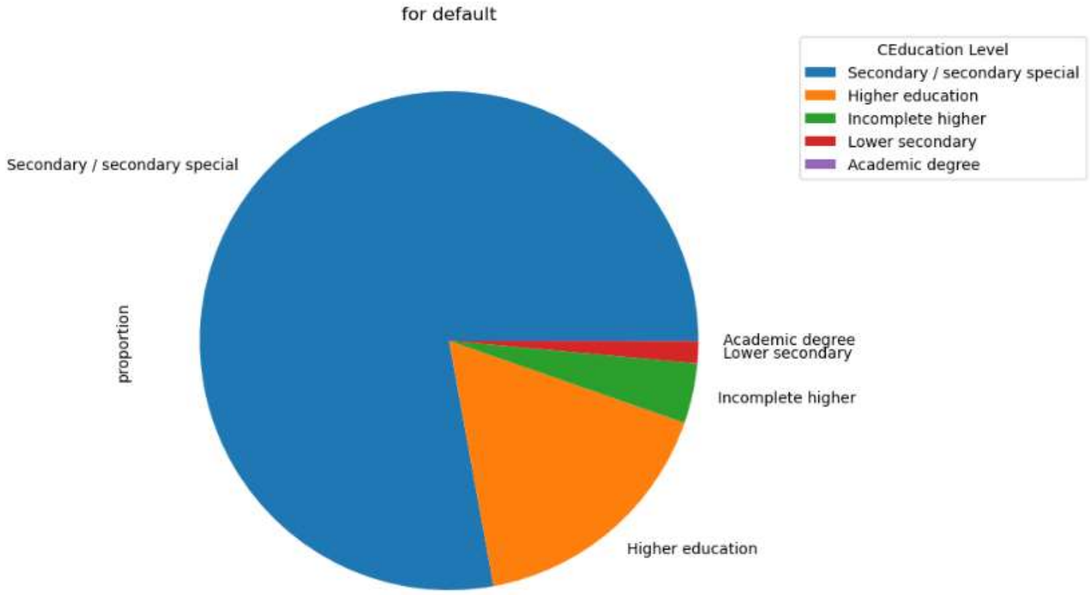
Target Variable Distribution

Shows non-defaulters dominate



Education Level Pie Chart

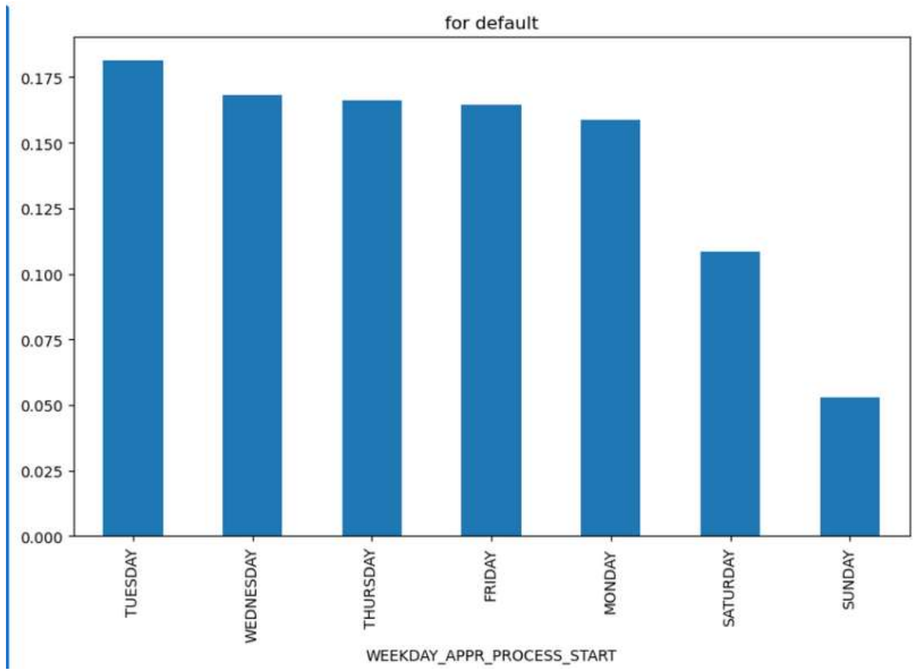
Shows secondary education is most common among defaulters



- 1 = defaulters (customers who failed to repay loans).
- 0 = non defaulter (customers who successfully repaid loans).

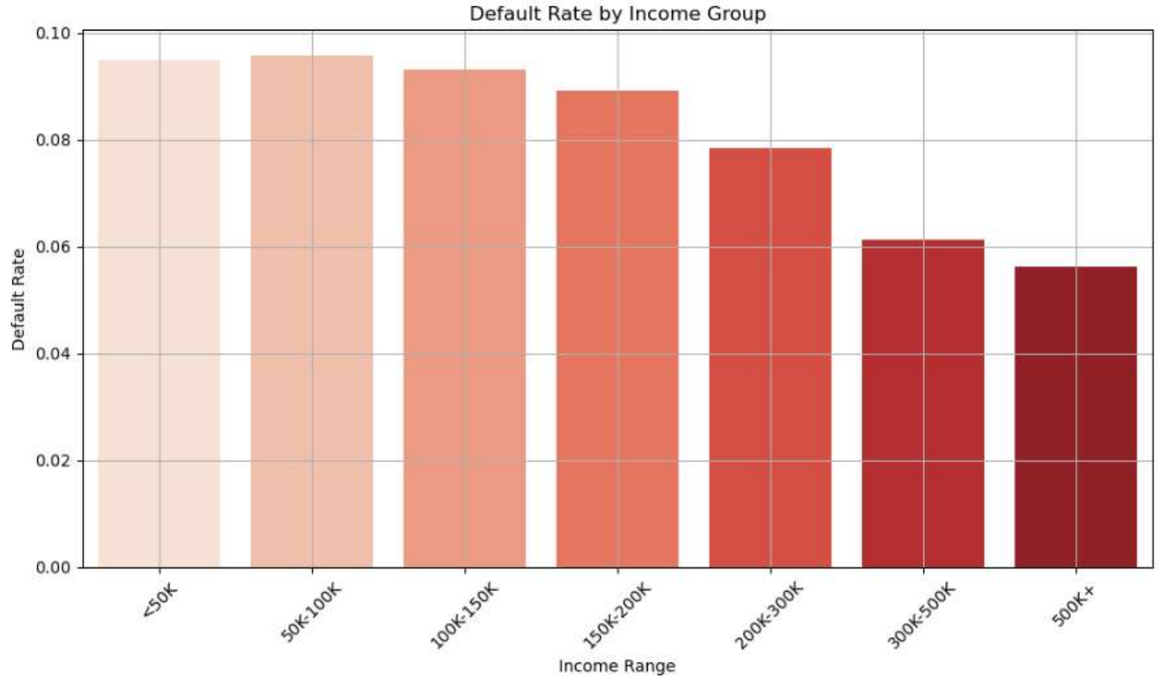
Application Timing Bar Chart

Shows activity highest on weekdays



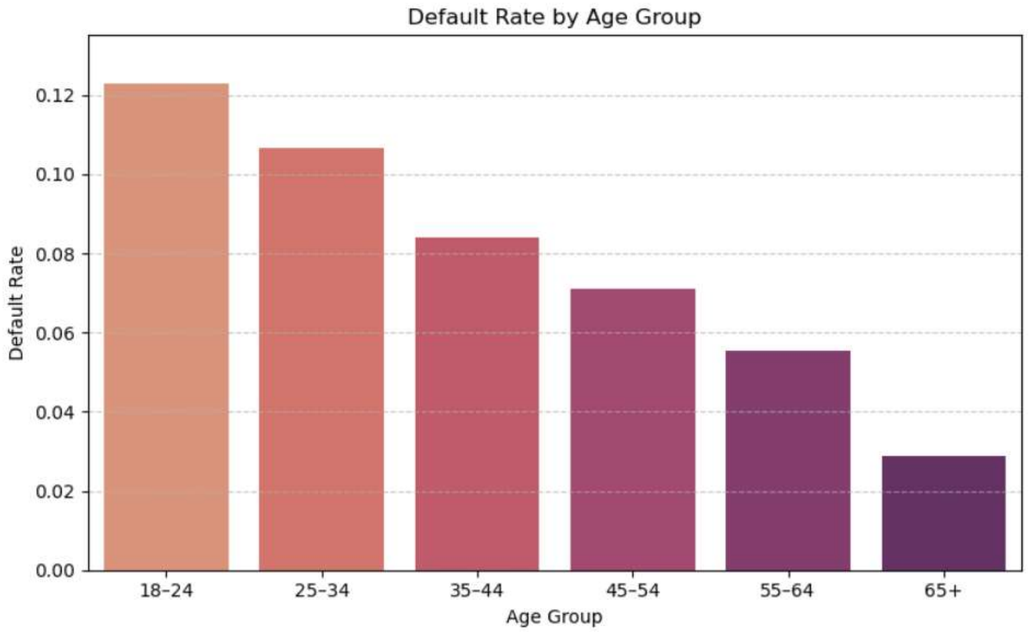
Income vs Default

Lower income is a significant risk indicator.



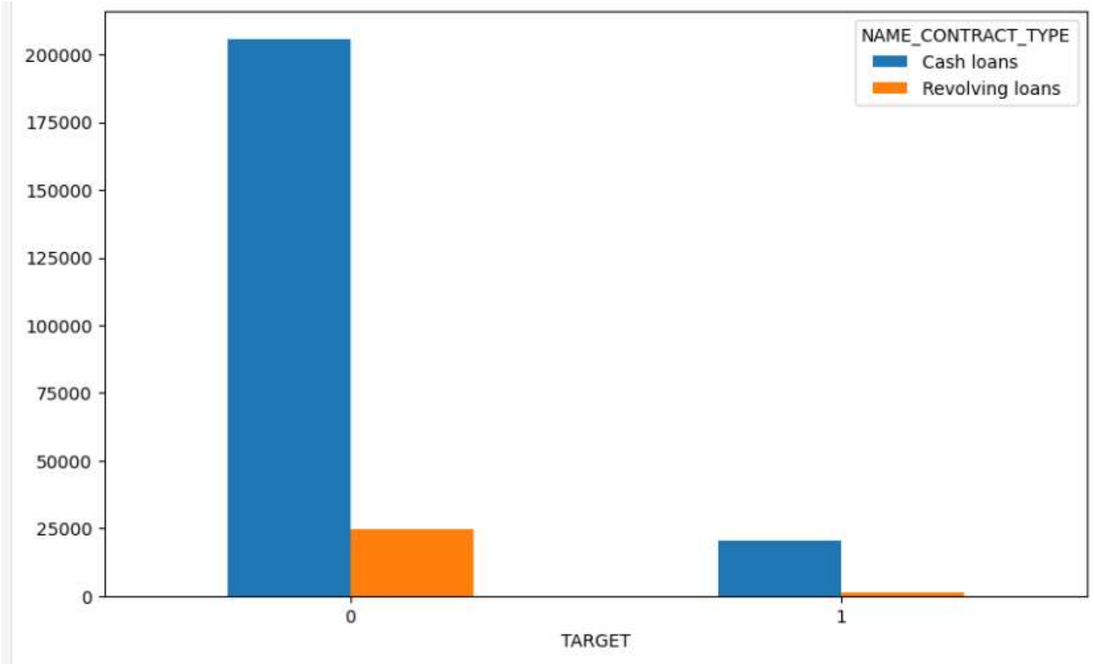
Age Distribution

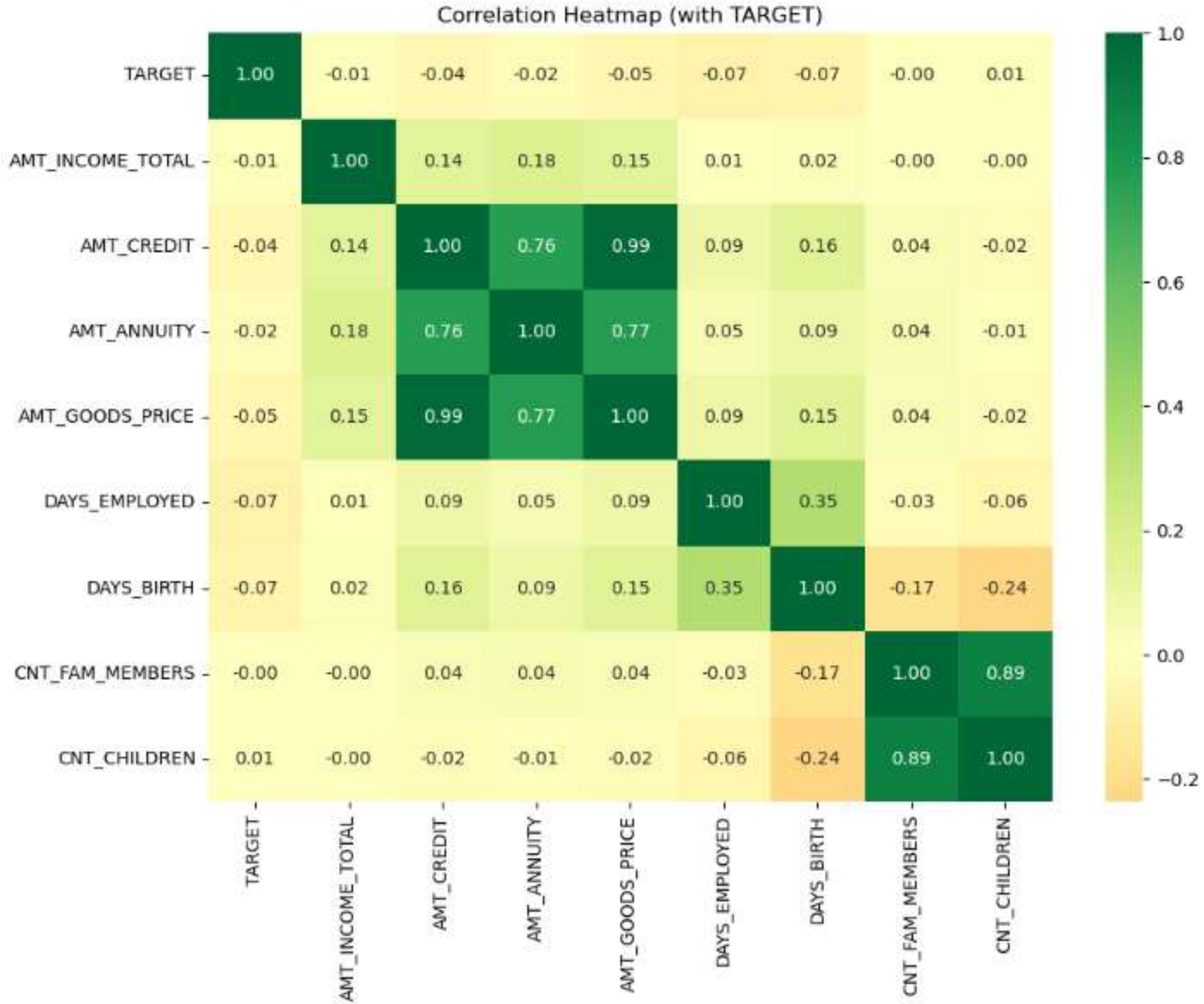
Proves younger customers are riskier



Loan Type vs Target

Proves cash loans = higher defaults



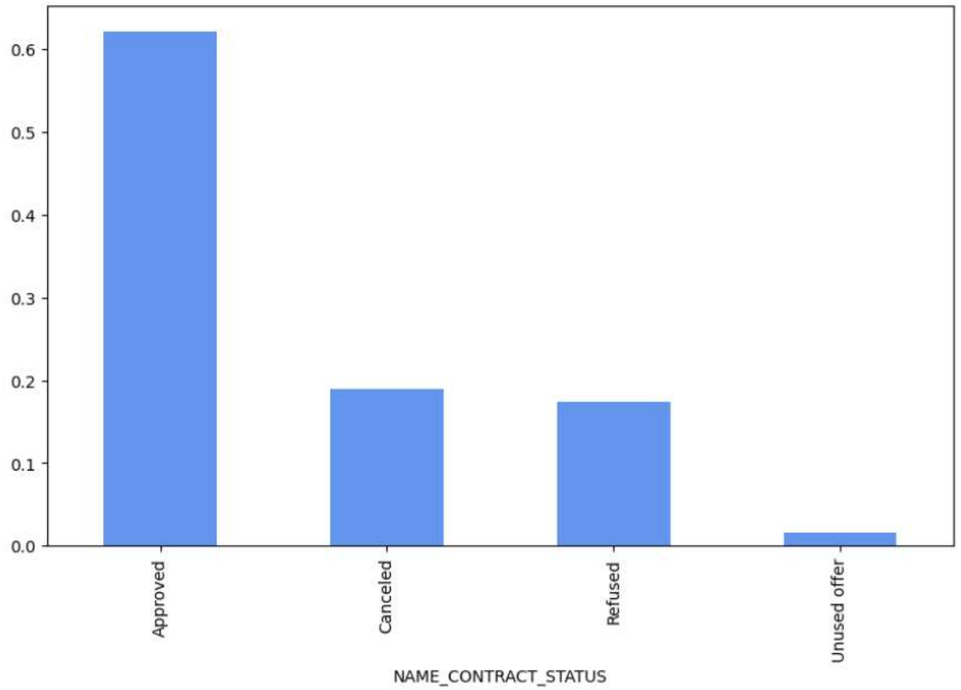


Correlation Heatmap

Shows strong links: income, credit, employment vs default

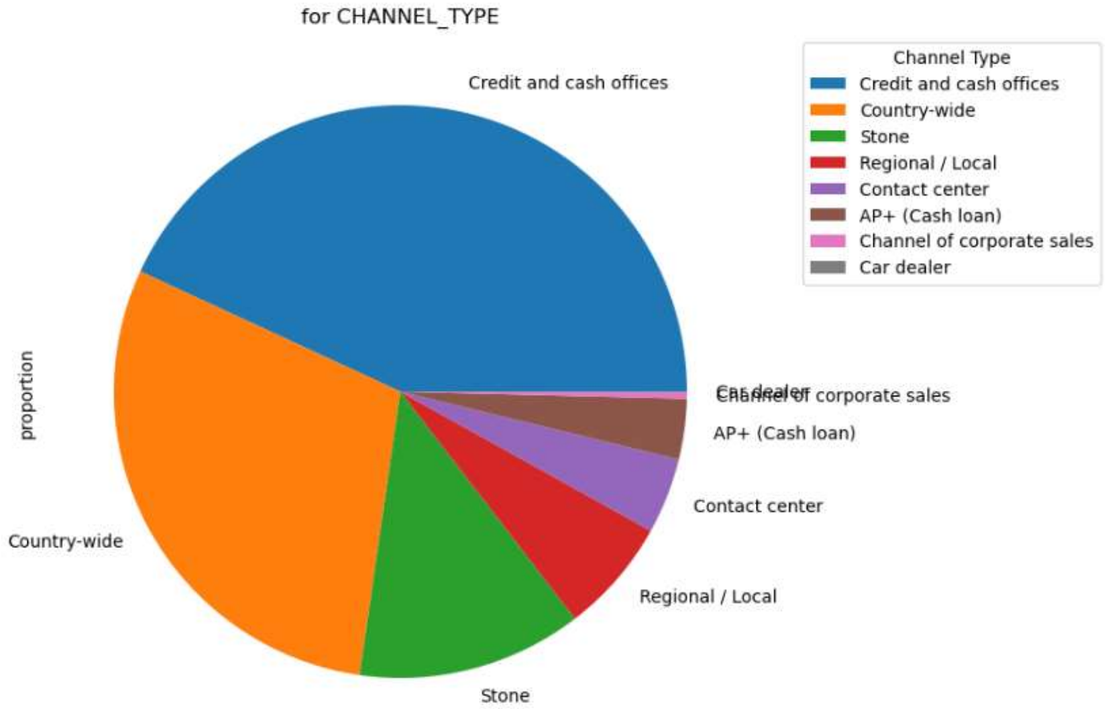
Loan Status Distribution

Shows previous refusals are significant



Channel Type

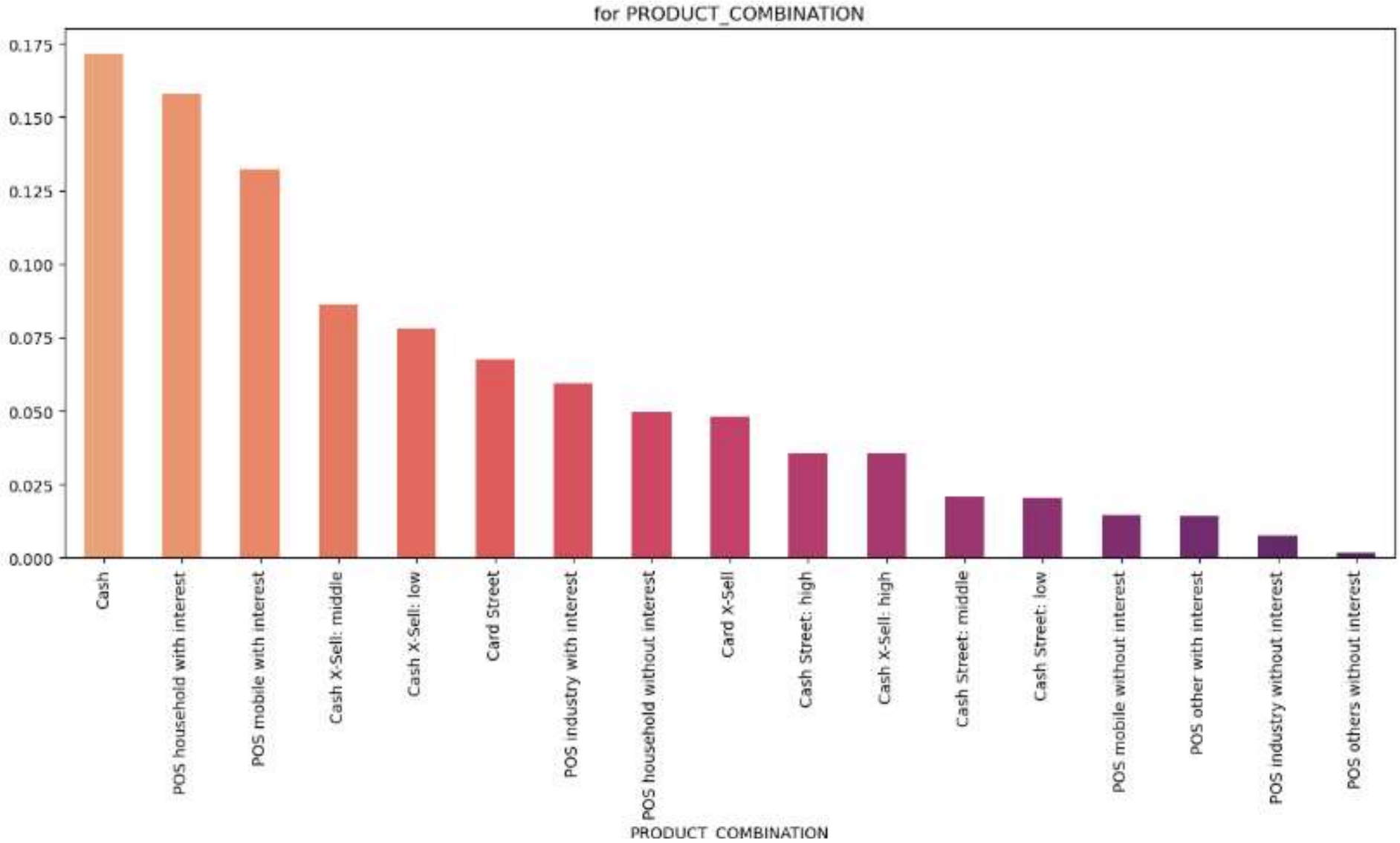
Supports how loans are processed



From previous_application.csv

Product Combination

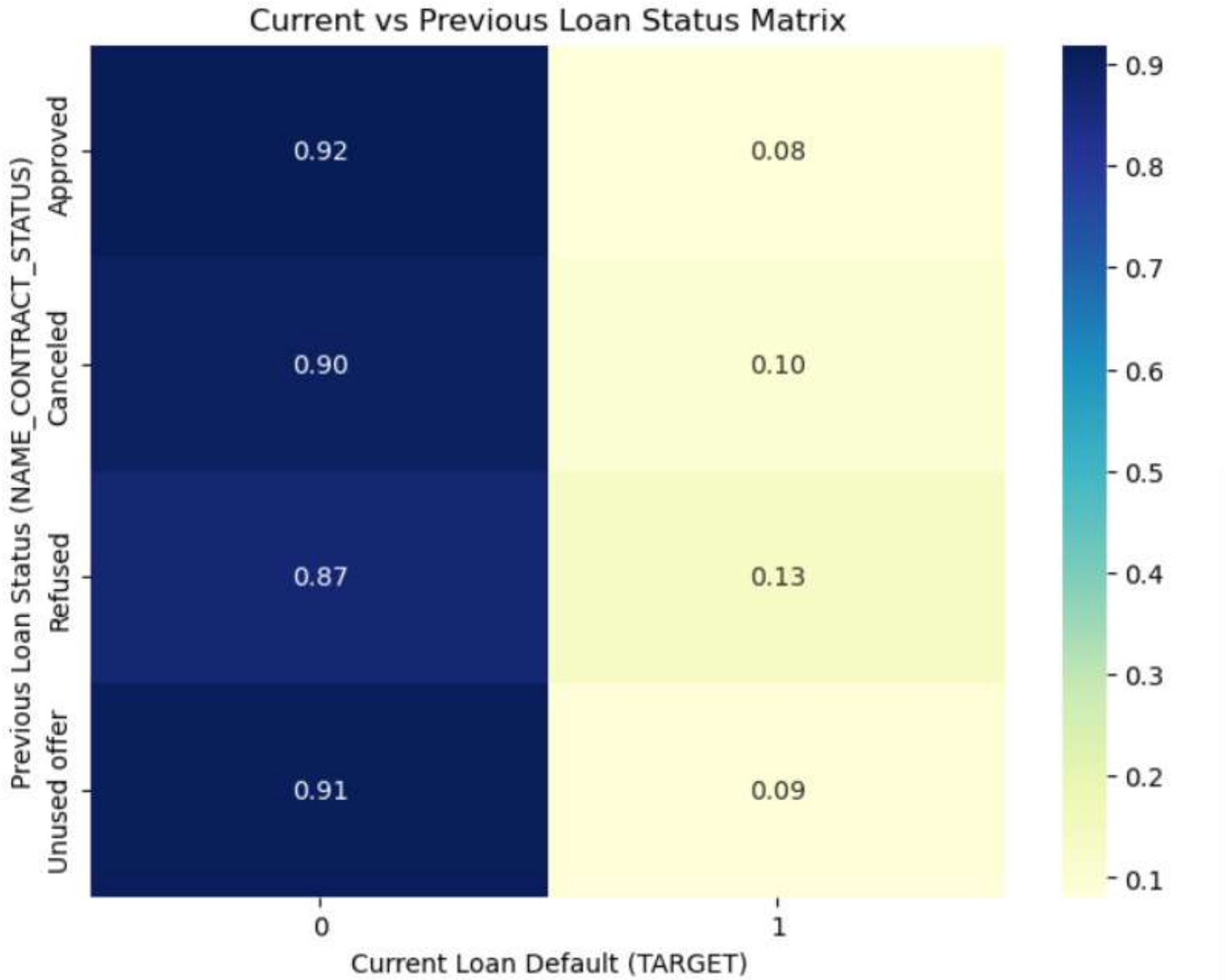
Spot patterns of simple vs bundled loans



From previous_application.csv

Current vs Previous Loan Status

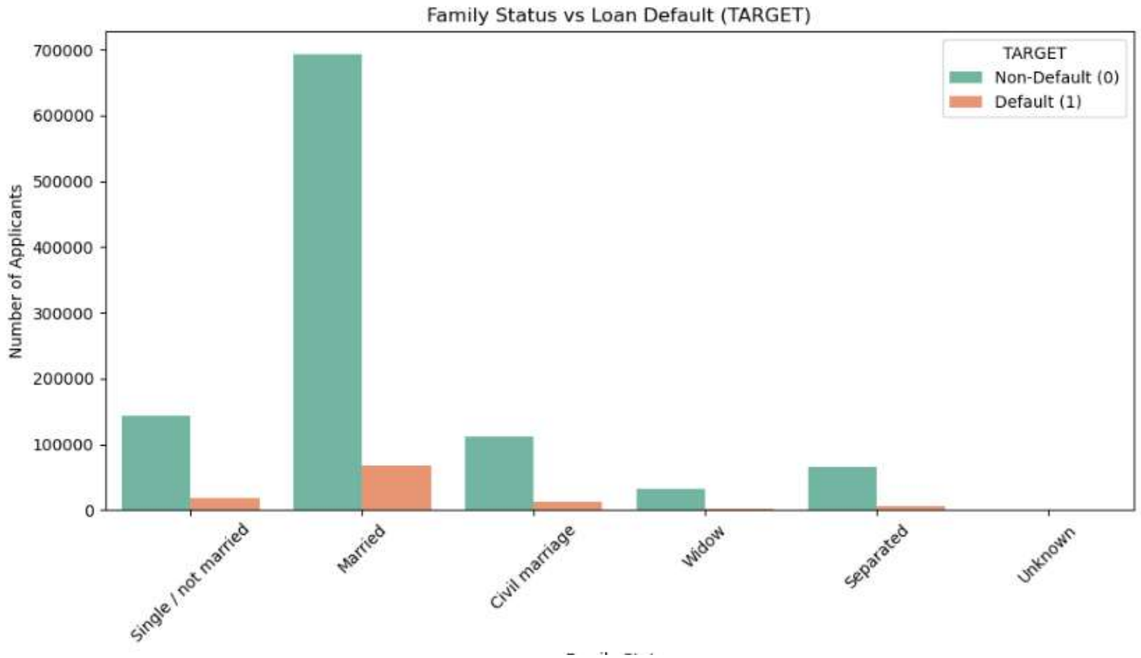
Links past refusals to current defaults



From Merged Dataset

Family Status vs Loan Status

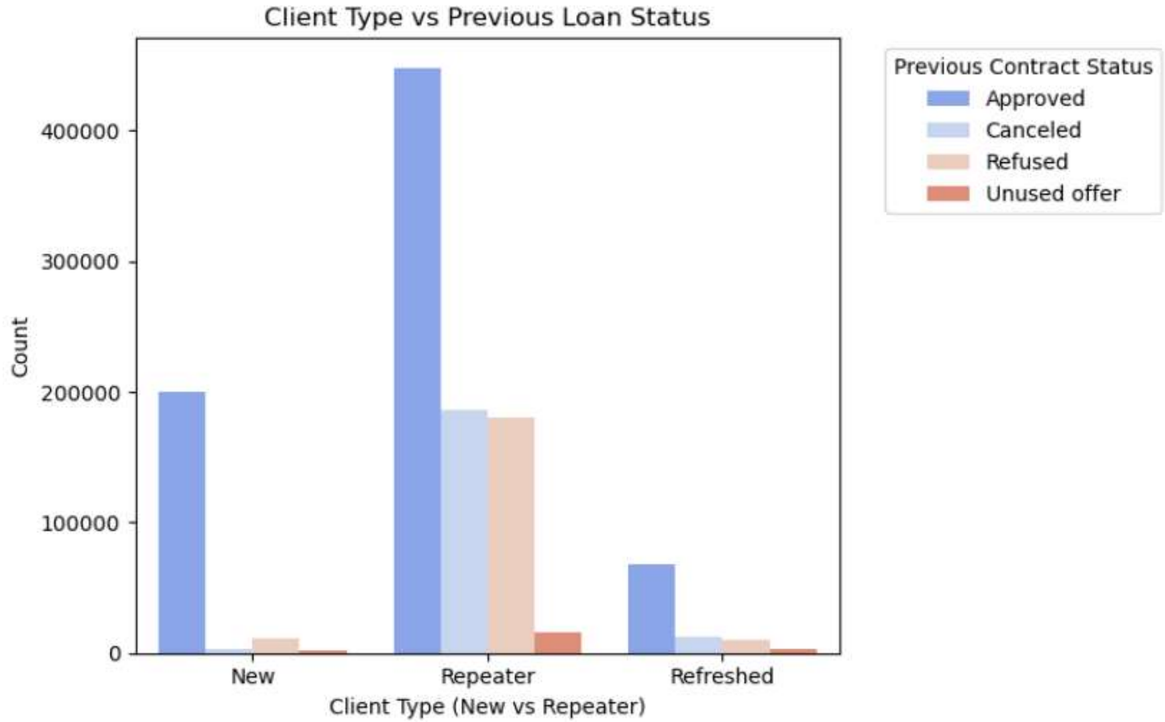
Adds socio-demographic dimension to profile



From Merged Dataset

Client Type vs Contract Status

Repeating clients more likely to be approved, safer



Conclusion

Customer Profile

Most customers are non-defaulters

Secondary education common among defaulters

Application activity highest on weekdays

Risk Factors

Lower-income customers show higher risk

Younger customers have higher default rates

Cash loans riskier than revolving loans

Strong correlations found between employment, income, and default rates.


THANK YOU