Predicting Loan Defaults to Minimize Risk
First Delivery

03/25/2025

Ibrahim Denis Fofanah
if57774n@pace.edu

Practical Data Science

MS in Data Science
Seidenberg School of Computer Science and Information Systems Pace University

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Key Business Takeaways and technical next steps

Executive summary

Lending institutions lose revenue when customers default on their loans. Without clear indicators, loan officers struggle to identify which applicants are most likely to default..

Solution:

- **Data Collection**: Used a dataset of 32,586 loan applications, including details such as loan amount, customer income, credit history length, and loan grade..
- **Exploratory Analysis:** Explored the relationship between borrower characteristics and default behavior.
- **Risk Insights:** Identified key drivers of default risk, such as short credit history and lower loan grades.
- **Outcome**: These insights will support more informed, risk-aware loan approval processes.

This approach helps lenders make data-driven decisions, reduce loan losses, and improve customer risk profiling.

Project Plan Recap

Deliverable	Due Date	Status
Data & EDA	03/25/2024	**Complete**
Methods, Findings, and Recommendations		**In Progress**
Final presentation		**Not Started**

Data

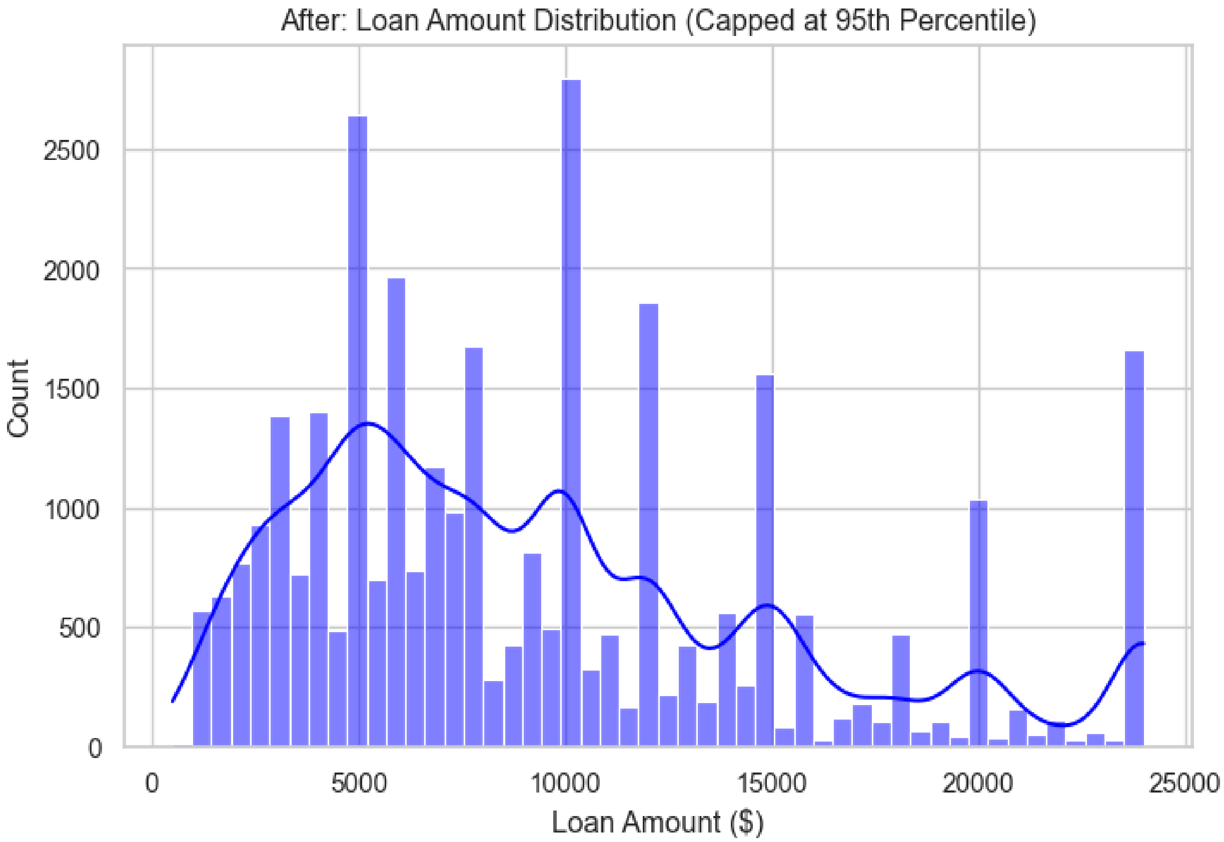
Data

- Data Overview:
    Data Source: Kaggle open dataset: Loan Default Prediction Dataset
    - **Dataset Url**: Loan-Dataset
    - **Sample size**: 32,586 rows, where each row represents a single customer loan application
    - **Time Period**: Time period not specified in the dataset — we assume it's collected over recent years by financial institutions
    - **Inclusion/Exclusion:** Retained only relevant features like loan amount, credit history, loan grade, default status, etc. and exclude customer id
    - **Clarifications:**
    - Missing values in loan amount were filled with median values
    - Extreme loan amounts were capped at the 95th percentile to minimize skewness

- Assumptions
    - We assume that loan grade was assigned based on the borrower’s creditworthiness
    - We also assume that credit history length is an important proxy for borrower trust
    - Since no dates were provided, we treat the data as a single snapshot of past loans

Exploratory Data Analysis

Loan Amount Distribution (Capped at 95th Percentile)



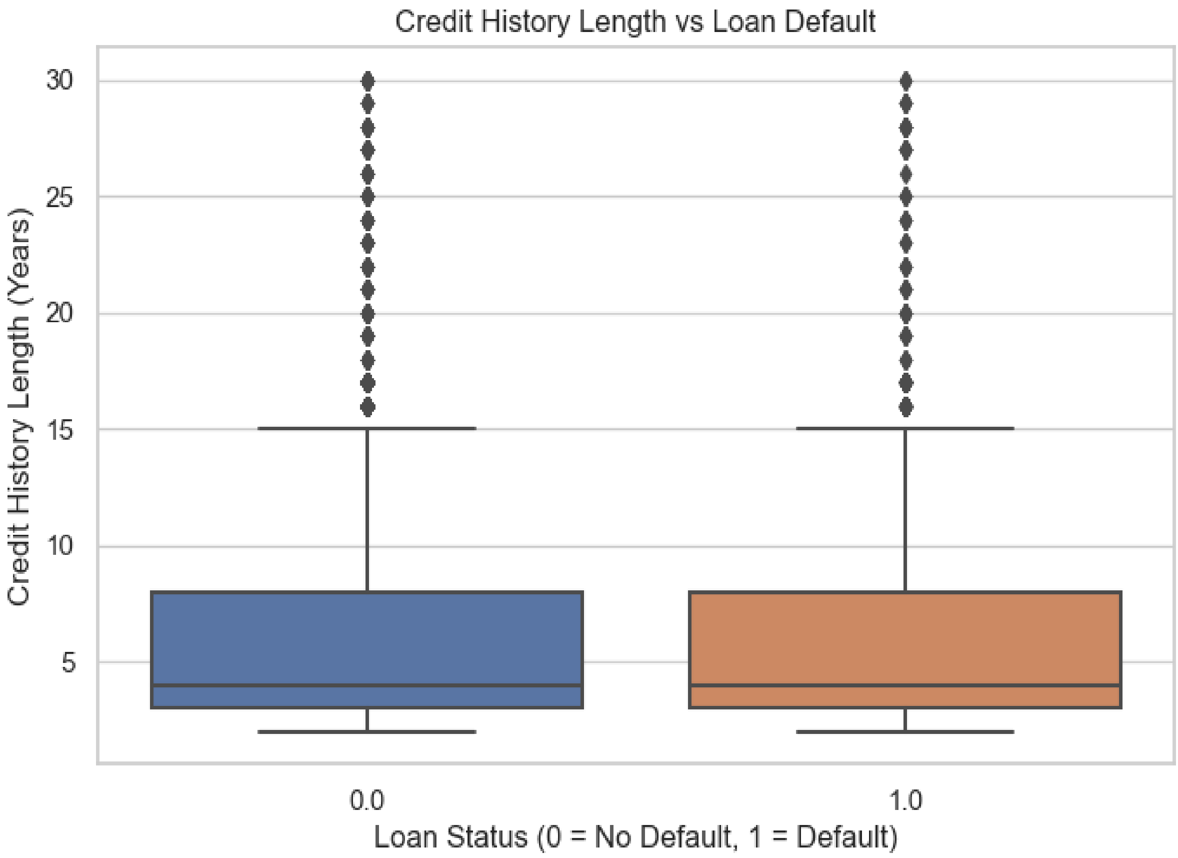
Key Takeaways

- Most borrowers request loans between $5,000 and $15,000
- Loan requests over $25,000 were capped to remove outliers
- Original data had extreme values up to $3.5M
- This cleaned view helps define what a 'normal loan' looks like

Data Notes

- Source: Kaggle Loan Default Dataset
- Sample Size: 32,586 rows (each represents one borrower)
- Time Period: Not provided – assumed to be recent
- Only loans capped at $25,000 are shown here for clarity

Credit History vs Loan Default



Key Takeaways

Borrowers who default tend to have **slightly shorter credit histories**, but the difference is **not very large**

• Both groups show a **similar distribution**, suggesting credit history **alone** may not strongly predict default

• However, **when combined with other features** (like income or loan grade), it could still contribute to identifying risk

Data Notes

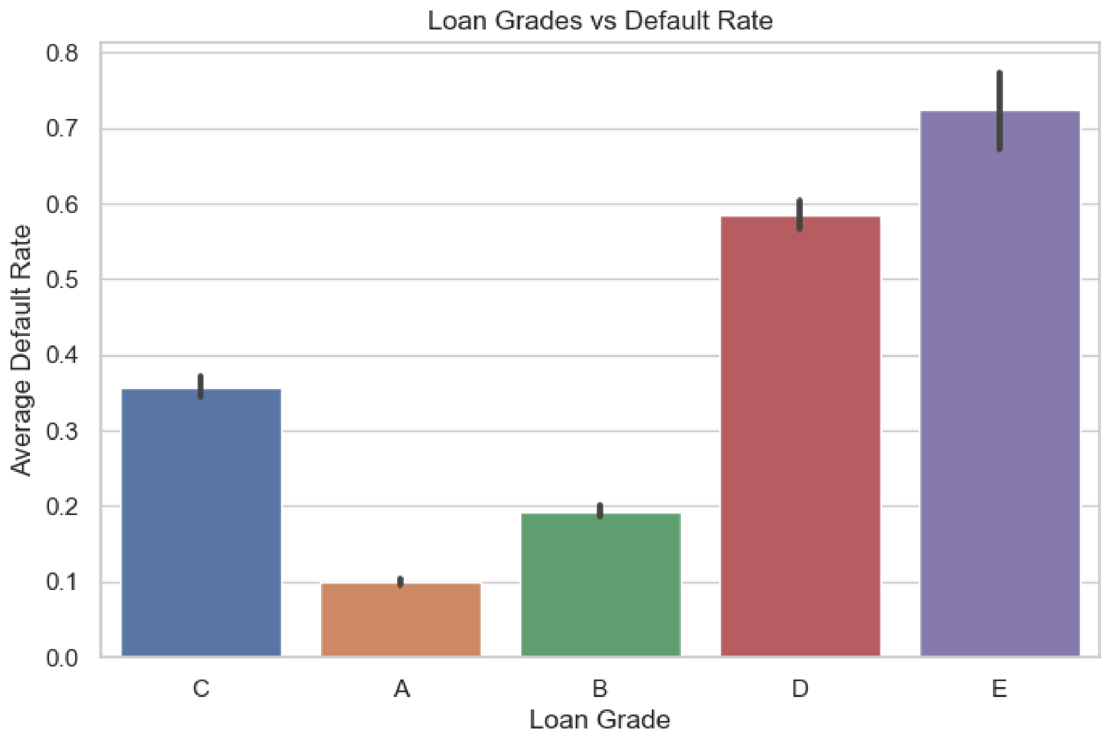
• Source: Kaggle Loan Default Dataset

• Sample Size: 32,586 rows

• Boxplot compares credit history in years for defaulted vs non-defaulted loans

We expected credit history length to show a bigger contrast between defaulters and non-defaulters. The similarity here reminds us that default prediction often depends on **a mix of features**, not just one.

Loan Grade vs Default Rate



Key Takeaways

• Loan grade is a strong signal of borrower risk
• Borrowers with Grade A have the lowest default rate (under 10%)
• Risk increases steadily from B to E — Grade E borrowers default over 70% of the time
Lenders should be more cautious with lower grades or adjust interest rates accordingly

Data Notes

• Source: Kaggle Loan Default Dataset
• Sample Size: 32,586 rows
• Chart shows average default rate by loan grade

Default rates grow sharply from Grade A to Grade E, making loan grade one of the most useful early warning signs in our data.

Key Business Takeaways and Technical Next Steps

Key Business Takeaways

- Most loans fall between $5,000 and $15,000 (some outliers were removed)
- Loan grade is a strong predictor of default — higher grades, less risk
- Credit history has subtle influence — may help when used with other features
- Data cleaning was essential to build valid and accurate insights

Technical Next Steps

- Engineer new features (e.g., buckets for credit history)
- Begin model training (logistic regression, decision trees)
- Evaluate models with accuracy, precision, recall metrics

Link to Git Repo for this Delivery

- https://github.com/Denis060/Loan_Default_Prediction