Predicting Loan Defaults to Minimize Risk Presentation

04/01/2025

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
- Modeling methods
- Findings
- Recommendations and technical next steps
- Appendix

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	04/01/2024	Complete
Final presentation		In Process

Data

Data

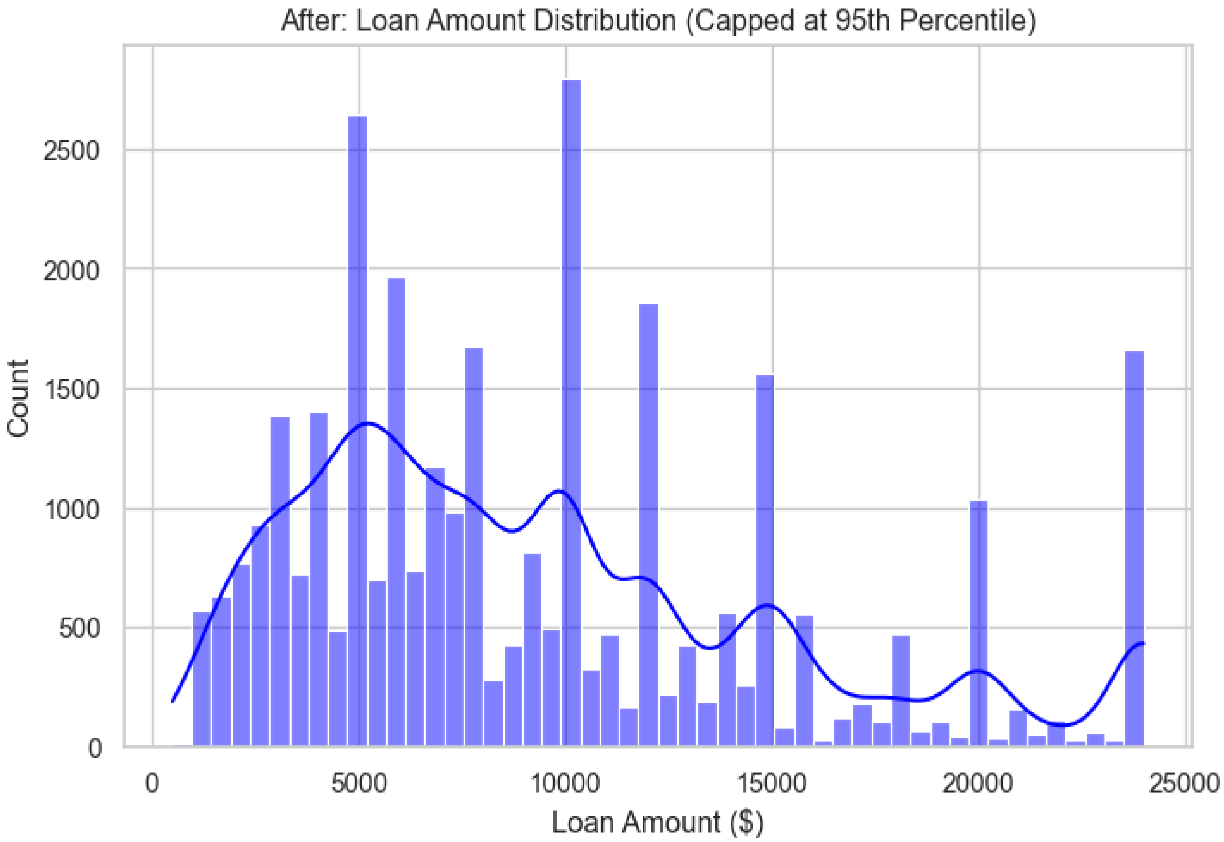
- **Data Overview:**

  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)

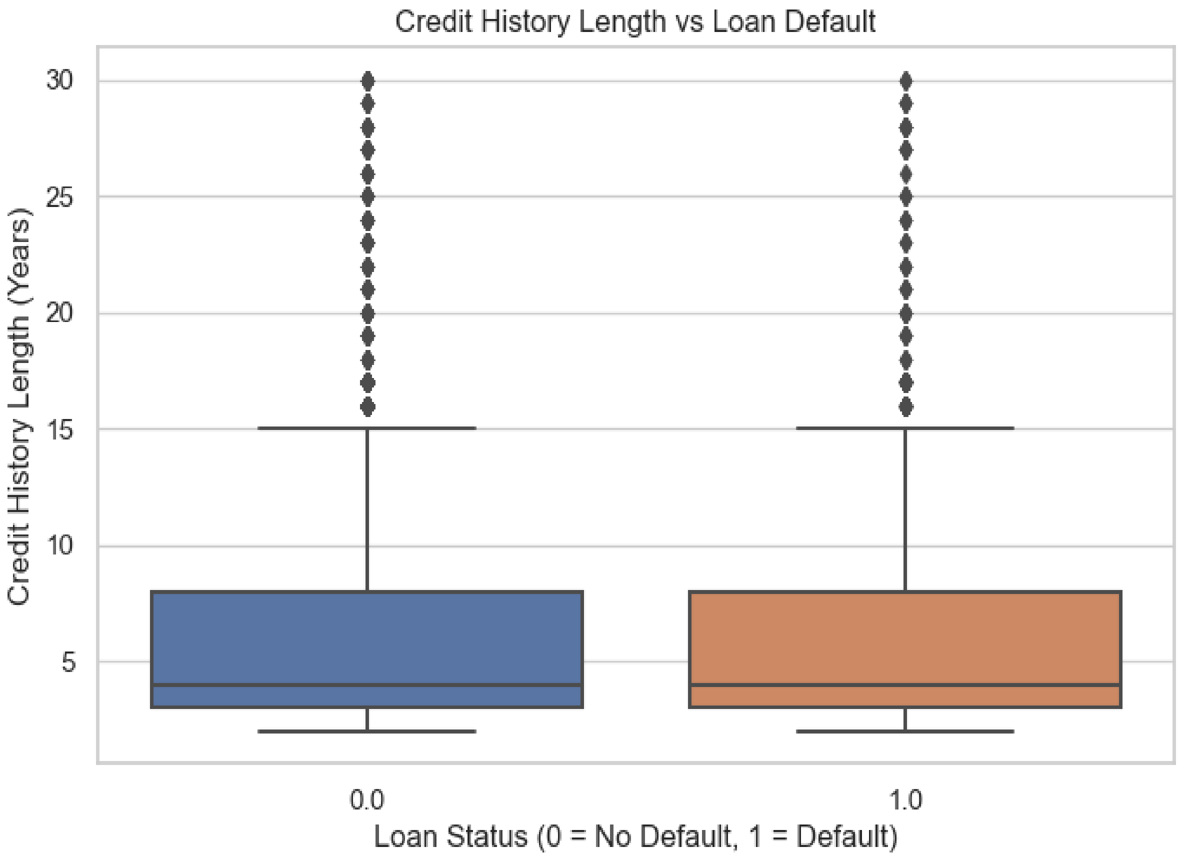


Key Takeaways

- 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

See Loan Amount Distribution before Capping

Credit History vs Loan Default



Key Takeaways

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

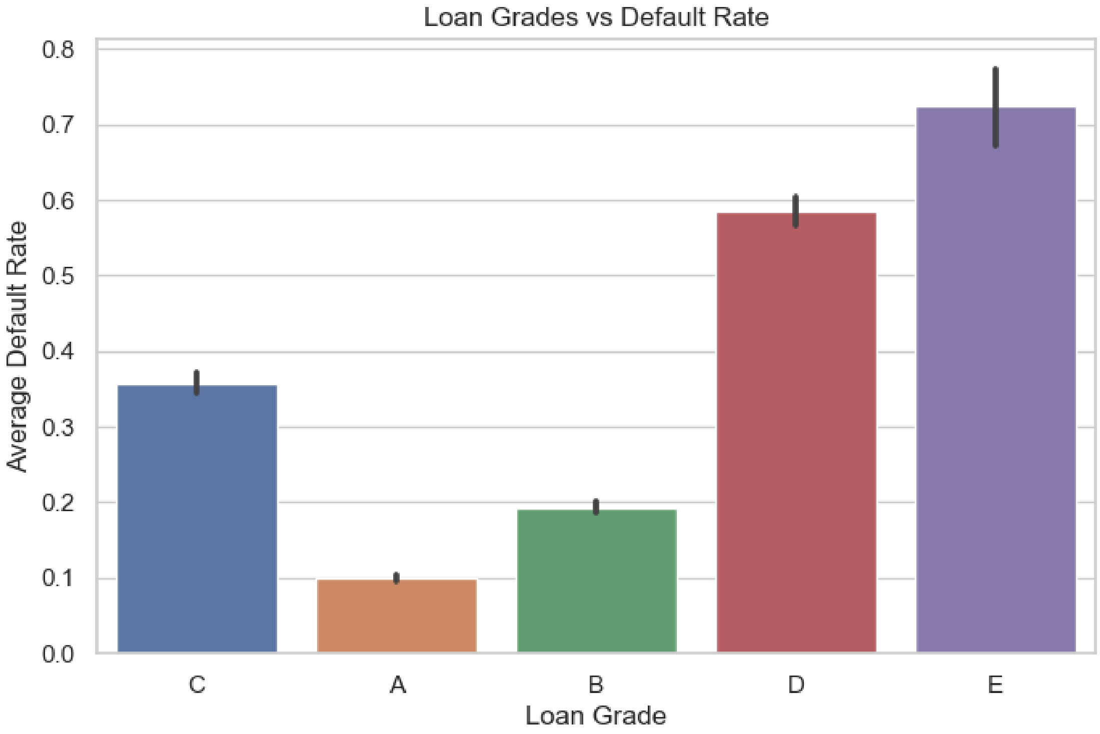
• Both groups show a **similar distribution**, suggesting credit history **alone** may not strongly predict default

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

• 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

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

Modeling methods

Modeling Methods

Outcome Variable:

Loan Default Status: We aim to predict whether a customer will default on their loan. Understanding this helps the organization reduce financial risk and make informed lending decisions.

Features Used with Hypotheses:

Below are the features used with some hypothesis on their behavior

1. **Customer Income:**
   Lower income may increase the risk of default due to limited repayment capacity.
2. **Loan Amount:**
   Higher loan amounts might lead to higher default risk.
3. **Employment Duration:**
   Longer employment suggests financial stability, possibly lowering default risk.
4. **Home Ownership:**
   Owning a home may indicate lower financial risk than renting.
5. **Loan Intent:**
   Some loan purposes (e.g., venture loans) may be riskier than others (e.g., home improvement).
6. **Credit History Length:**
   A longer credit history typically reflects more financial experience and better creditworthiness.

Chosen Model Type: Logistic Regression

What is Logistic Regression?

Logistic Regression is a statistical model that helps predict **one of two outcomes**, like **default** vs **no default**. Instead of drawing a line like in linear regression, it draws a boundary that separates the categories..

Why Logistic Regression?

We chose Logistic Regression because:

- It's simple and interpretable
- Ideal for **binary outcomes** like predicting whether a customer will **default on their loan**
- Helps us understand **how each feature affects the risk of default**
- Gives us **probability scores**, allowing risk-based decision-making

This model helps financial institutions quickly assess risk based on customer features like income, loan amount, and credit history.

- **How it works**: Logistic regression assigns weights to each feature and combines them to estimate the **probability** of default. If the probability is greater than 0.5, it predicts **default**. Otherwise, it predicts **no default**..
- **Example**: If low income is strongly associated with default, the model will learn this pattern and flag low-income customers as higher risk.

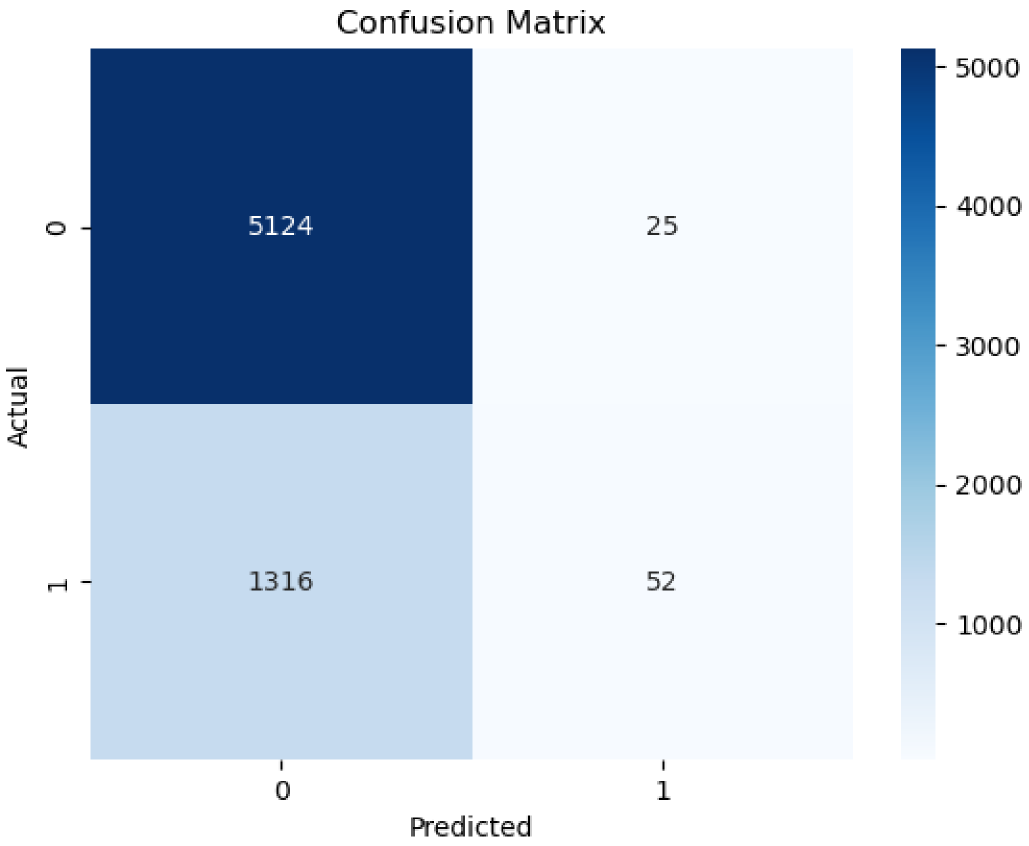
Findings

Results and Interpretation

Model Performance

- **Accuracy:** 79%
  The model correctly predicted 79% of the loan statuses (Default or No Default). This shows it performs well overall, especially in identifying borrowers who will repay.
- **Precision (Default class):** 68%
  When the model predicts someone will default, it's correct 68% of the time.
- **Recall (Default class):** 4%
  The model struggled to identify actual defaulters, catching only 4%. This is due to class imbalance, meaning there are far more non-defaulters than defaulters in the dataset.
- **F1-Score (Default class):** 7%
  A balance of precision and recall – in this case, low because of recall.

Results and Interpretation



- **True Negatives (5124):** Correctly predicted No Default
- **False Positives (25):** Predicted default, but actually No Default
- **False Negatives (1316):** Predicted No Default, but actually defaulted
- **True Positives (52):** Correctly predicted Default

Interpretation:

- The model performs well for predicting borrowers who won't default.
- However, it misses most of the actual defaulters, which is risky in real-world lending.
- Future steps may involve techniques to handle imbalance, like resampling or advanced models.

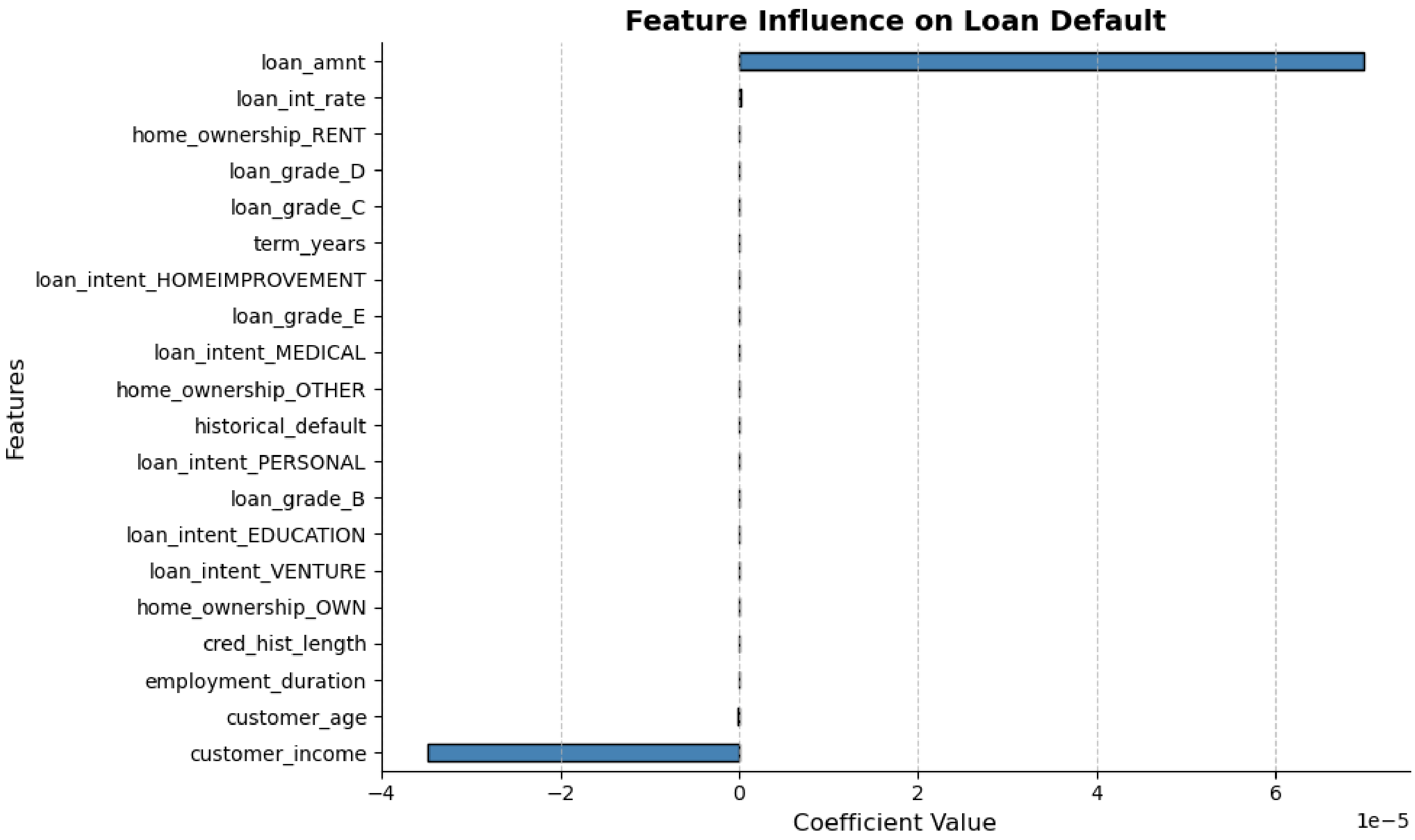
Feature Analysis and Business Insights

Each feature shows how likely it is to influence whether a customer will default on their loan

- **Customer Income (↓):** Higher income significantly lowers the chances of default. This suggests customers with better financial stability are less risky.
- **Loan Amount (↑):** Larger loan amounts slightly increase the risk of default. Bigger loans may become harder to repay, especially for lower-income borrowers.
- **Interest Rate (↑):** Higher interest rates are linked to more defaults. Borrowers with higher rates may already be considered riskier or may find repayments harder over time.
- **Loan Grade E & D (↑):** These grades are associated with higher default risk. These loans likely represent higher-risk segments already flagged during underwriting.
- **Employment Duration (↓):** Longer employment is weakly linked to reduced default risk. Stable employment can indicate financial consistency.

- Refer to this link for a graphical representation of the feature coefficients.

Bar Chart Showing the Feature Coefficients



Recommendations & Data Science Next Steps

Recommendations

1. **Focus on Applicants with Higher Income**
2. **Finding**: The model shoed customer income had a strong negative influence on default(higher income= less likely to default)

- **Connection to Business Problem**: Lending to customers with more stable income reduces the risk of default, improving the overall profitability of the loan portfolio.
- **Recommendation**: Priotized applicants with higher income buckets and consider income verification as a key screening metric.

2. **Reevaluate Riskier Loan Grades**

- **Finding**: Loan grades like D and E had significantly higher default rates.
- **Connection to Business Problem**Approving high risk loans without adjustments can lead to increase losses and affect business sustainability
- **Recommendation**: Adjust Approval criteria for lower loan grades or attached higher interest rates/risk based pricing models for such applicants

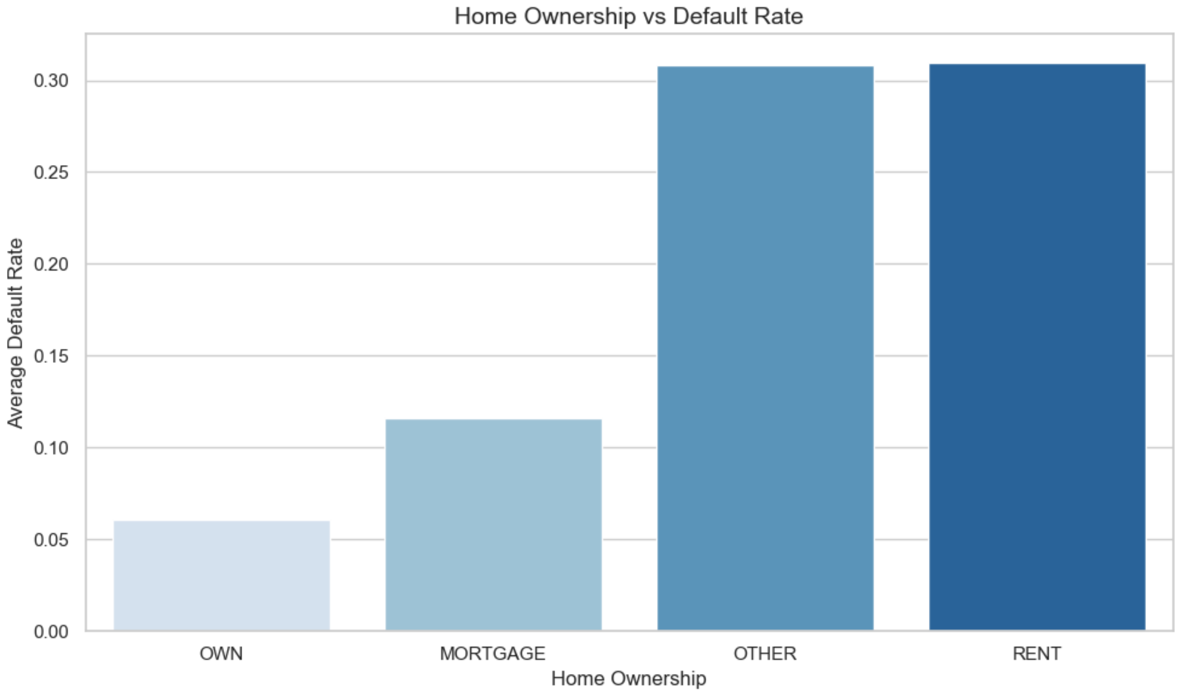
Data Science Next Steps

These are some technical directions the data science team could explore to enchance model accuracy and gain deeper insights:

1. **Try More Advanced Models**: Experiment with ensemble methods like Random Foerest, Gradient Boosting or **XGBoost** to better capture non linear relations between features and defaults. These models could potentially improve recall for default cases, which was low in the logistic regression model.
2. **Address Class Imbalance:** The dataset shows a high class imbalance, with far fewer default cases than non-defaults. Techniques such as **SMOTE**(Synthetic Minority Oversampling technique) or class weight adjustment could help the model focus more on detecting defaults accurately.
3. **Incorporating External Data:** Include additional socioeconomic data like **credit scores**, **employment sectors**, or **regional economic indicators** to improve predictive power. This could help the model understand **why** certain customers default.

Appendix

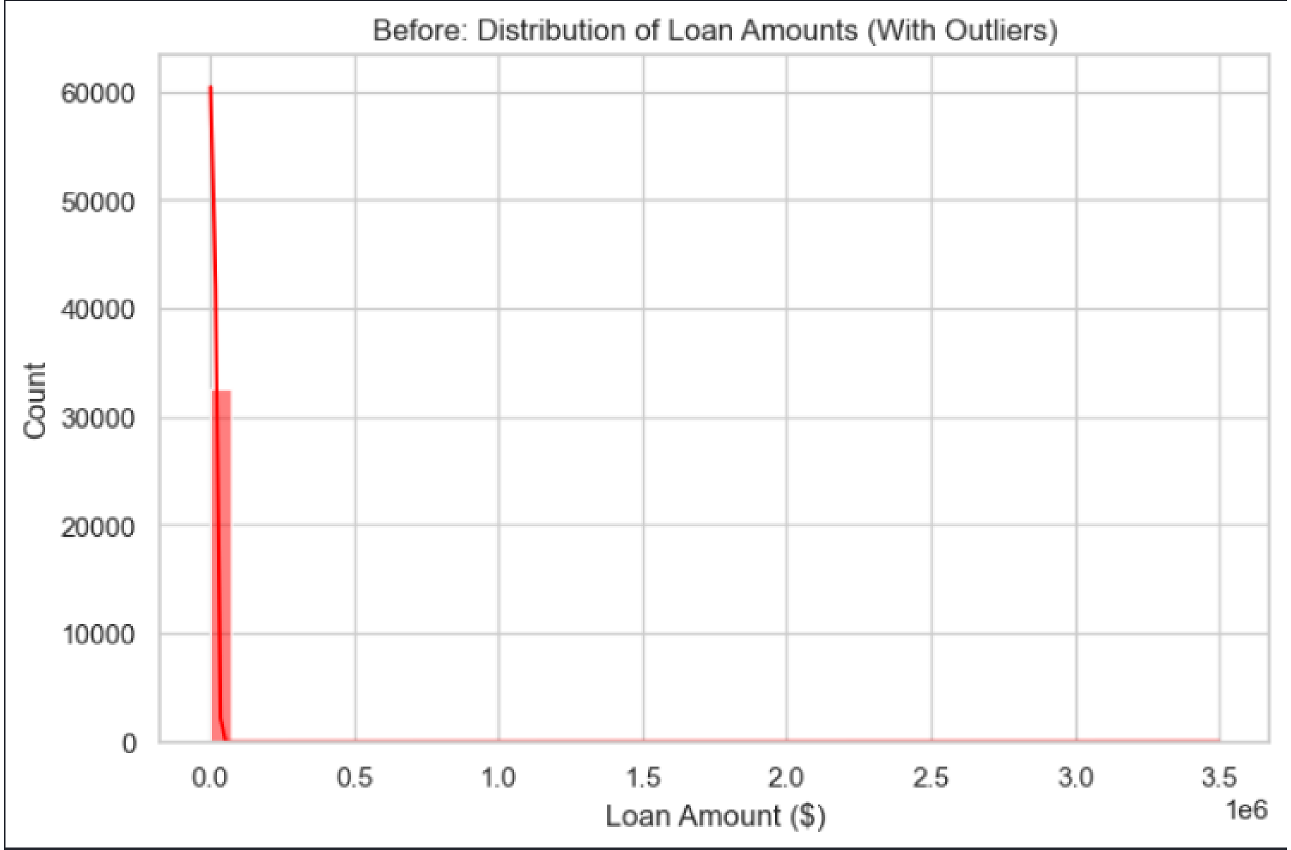
Home Ownership vs Default Rate



Borrowers who **rent** or fall under the **"other"** category exhibit significantly **higher default rates** compared to those who **own** their homes. This suggests that **homeownership status** is a strong predictor of default risk. Renters may face more financial instability, making them more likely to default on loans.

This insight can help lending institutions assess **risk levels** and potentially tailor lending strategies or interest rates based on homeownership status.

Loan Amount Distribution Before Cleaning



This chart displays the original distribution of loan amounts in the dataset. A few extremely high loan values (visible on the far right) skewed the distribution, making most loans appear clustered on the left.

These outliers were capped at the 95th percentile to improve model performance and visual clarity in analysis.

Return to: Loan Amount Distribution After Cleaning

Classification Report

Classification Report:

	precision	recall	f1-score	support
0.0	0.80	1.00	0.88	5149
1.0	0.68	0.04	0.07	1368
accuracy			0.79	6517
macro avg	0.74	0.52	0.48	6517
weighted avg	0.77	0.79	0.71	6517

The model performs well for **non-default** predictions (Class 0) with high precision and recall.
However, it struggles to identify **defaults** (Class 1), with a **very low recall (0.04)**, meaning most defaulters are missed.
The **macro average F1-score of 0.48** suggests **imbalanced performance** across classes.
This highlights a need for improving recall on the minority class (ults), possibly using techniques like resampling or model tuning.

See Confusion Matrix

Project Code Repository

All code for data cleaning, model training, and analysis can be accessed here:

https://github.com/Denis060/Loan_Default_Prediction