

# Credit Risk Intelligence

## Analytical Insights and Predictive Solutions

Project-Based Virtual Internship:  
**Data Scientist at ID/X Partners x Rakamin Academy**

---

Presented by:  
**Az-Zukhrufu Fi Silmi Suwondo**



## Company Profile

# About Company

id/x partners is an **Indonesian consulting firm** specializing in **Data Analytics, Decisioning, and RegTech solutions** since 2006. With 18+ years of experience, we have helped 75+ institutions, including top banks, multifinance companies, fintechs, and insurers, leverage advanced analytics and AI-based solutions to solve challenges in **digital lending, risk management, financial crime prevention, and regulatory compliance**.

This proven track record is recognized through **multiple industry awards**, including CIO Advisor's Top 10 APAC Data Analytics Consulting (2019) and four consecutive SAS Partner Appreciation Awards (2022-2025).



# Project Overviews

## Problem

Lending platforms face **significant credit risk** from **borrower defaults**, leading to **substantial financial losses** without effective risk assessment.

## Goal

Develop a **predictive credit risk model** achieving **0.70+ AUC-ROC** score to classify borrowers into risk tiers, enabling data-driven lending decisions and portfolio risk optimization.

## Objectives



### Exploratory Data Analysis

Identify key factors correlated with default risk through comprehensive data exploration.



### Model Development

Build and optimize a predictive model to estimate default probability using feature engineering and algorithm comparison.



### Risk Tiering System

Convert probability scores into actionable risk tiers with recommendations for each tier.



### Business Impact Assessment

Quantify potential loss reduction and provide insights for portfolio management.

# Dataset Overviews

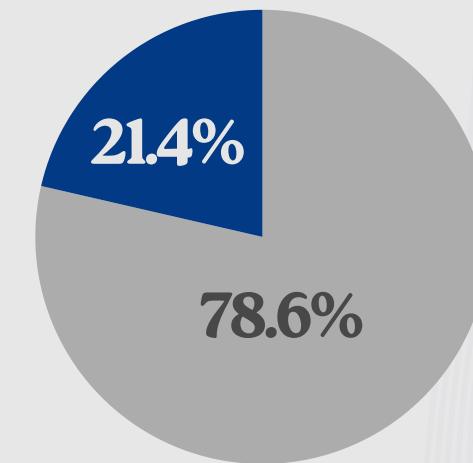
## Dataset Size



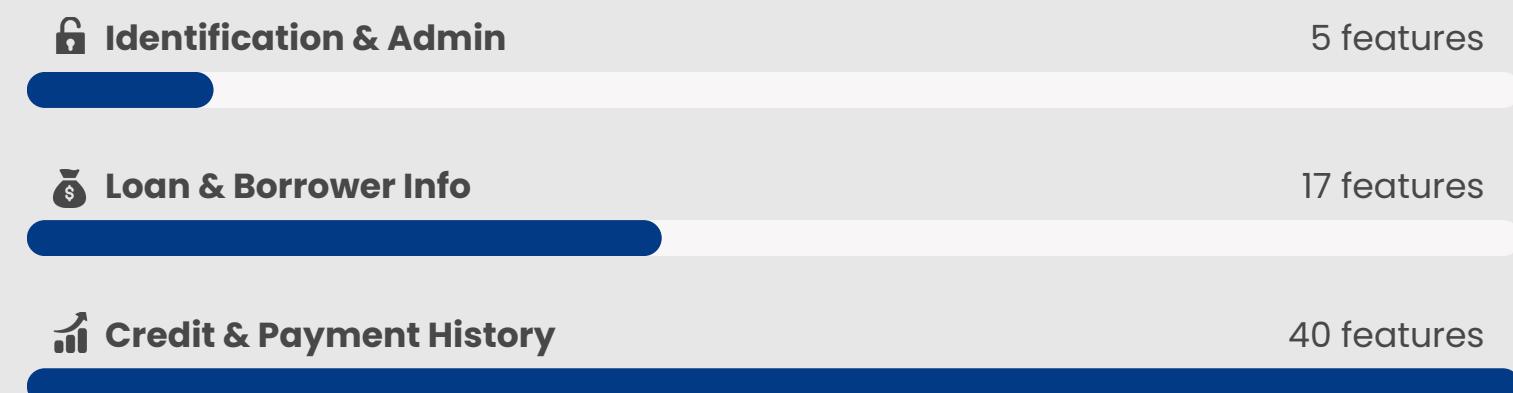
## Target Distribution

- Good Loans**  
(Fully Paid)
- Bad Loans**  
(Charged Off, Default, Late >31 days)

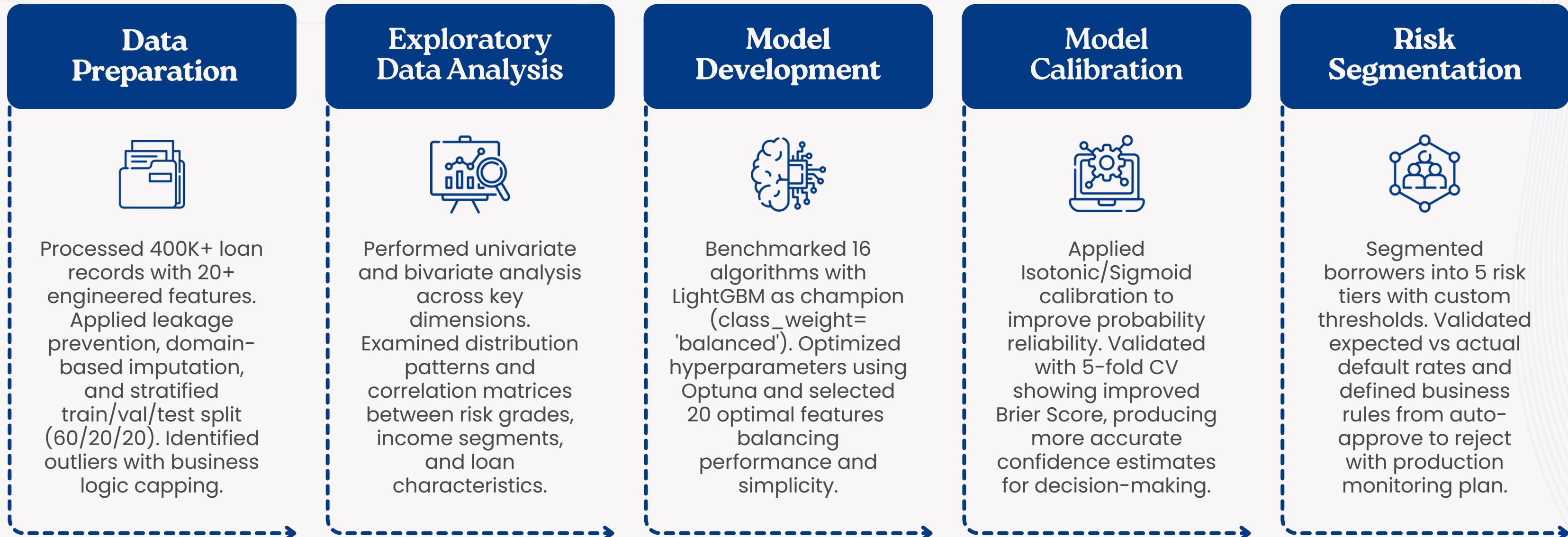
\*Includes policy-exception loans classified by final outcome  
\*Excludes current loans and in-progress statuses



## Feature Categories



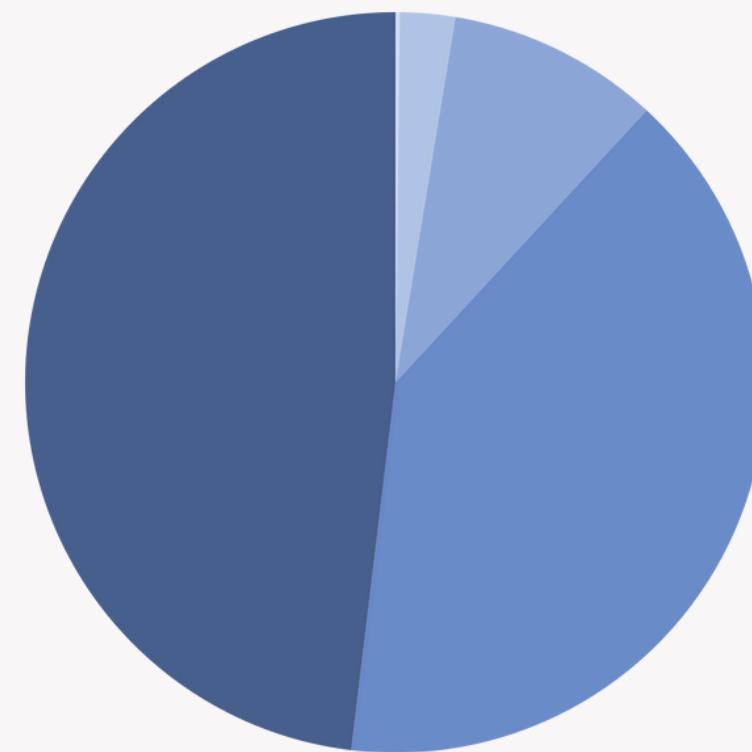
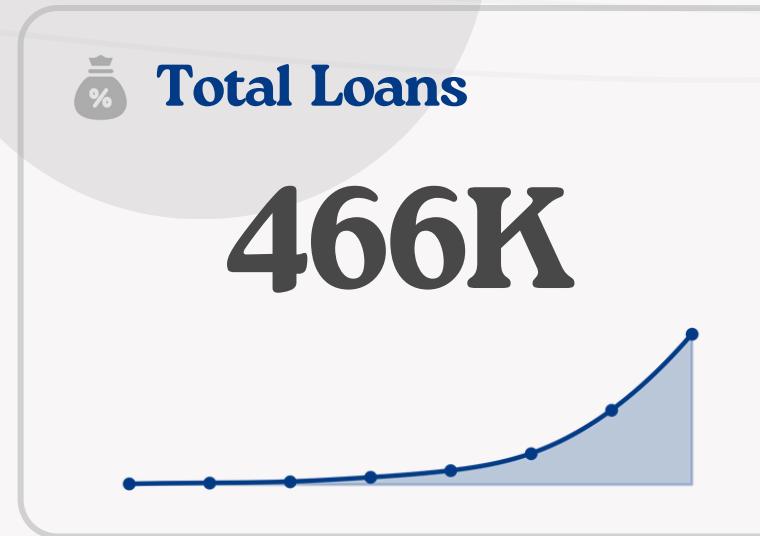
# Credit Risk Analytics Workflow



Industry-standard workflow **adapted from CRISP-DM framework**, tailored for credit risk modeling: **comprehensive data preparation** with leakage prevention, **exploratory analysis** revealing key risk indicators, **systematic model development** achieving 0.70+ AUC, **probability calibration** for decision reliability, and **5-tier risk segmentation** with production monitoring, delivering **end-to-end credit risk intelligence**.



# Portfolio Composition



-  Current (**48.1%**)
-  Fully Paid (**40.0%**)
-  Charged Off (**9.3%**)
-  Late/Delinquent (**2.4%**)
-  Default (**0.2%**)

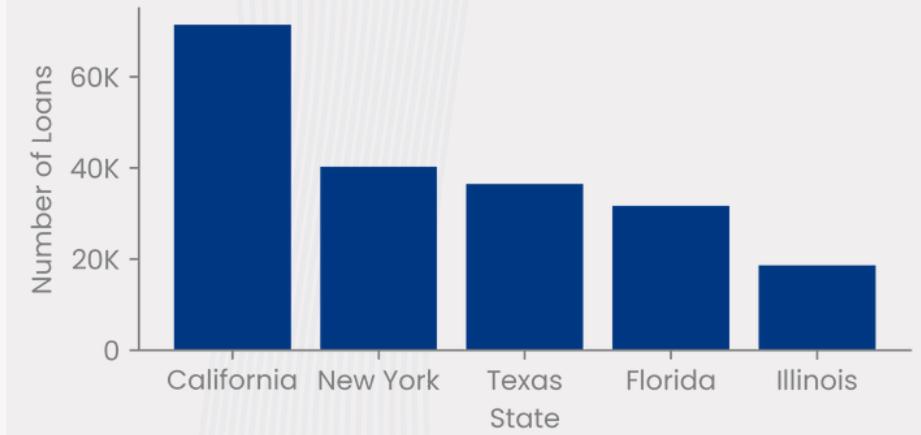
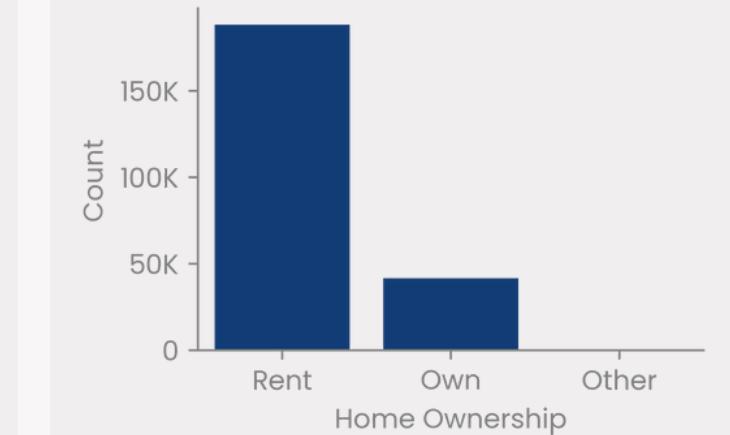
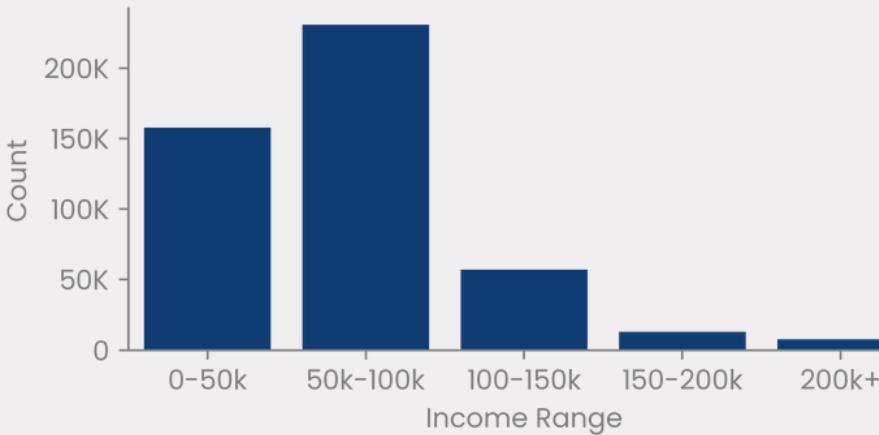
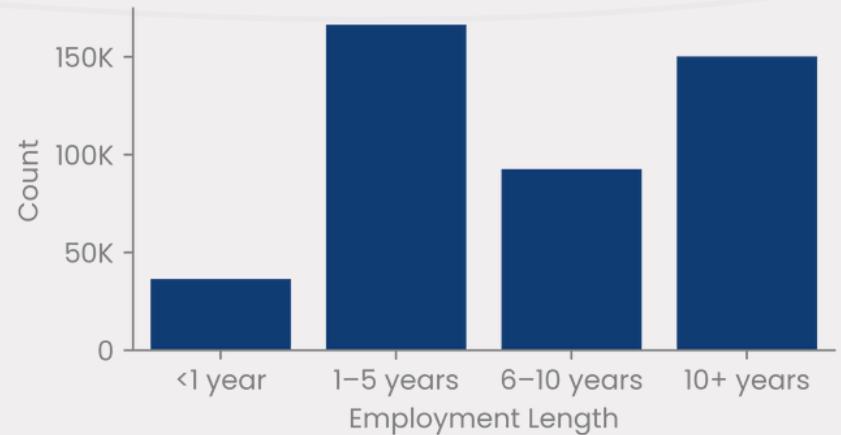
 Actionable Item

Continue monitoring late/delinquent accounts to prevent potential defaults and maintain portfolio health above 85%.

The portfolio demonstrates **strong performance** with 87.7% health rate and controlled risk levels. Current loans represent 48.1% of the portfolio, while fully paid loans account for 40%, indicating steady repayment activity. The minimal default rate of 0.2% reflects **effective credit assessment and monitoring practices**.



# Customer Profiling



## Experienced Workforce Dominance

37.4% have 1-5 years employment, followed by 33.7% with 10+ years. Only 8.2% are in their first year, indicating overall employment stability.



## Middle-Income Concentration

51.9% earn \$50k-100k, representing the core segment. Only 4.4% earn above \$150k, showing a middle-class customer base.



## Predominantly Renters

81.9% are renters (188,473), while just 18.1% own homes. This signals lower asset ownership and potentially higher credit risk.



## Geographic Concentration

Top 5 states (California, New York, Texas, Florida, and Illinois) represent the majority, with California leading at 71,450 customers.

Total Customers  
**466,285**

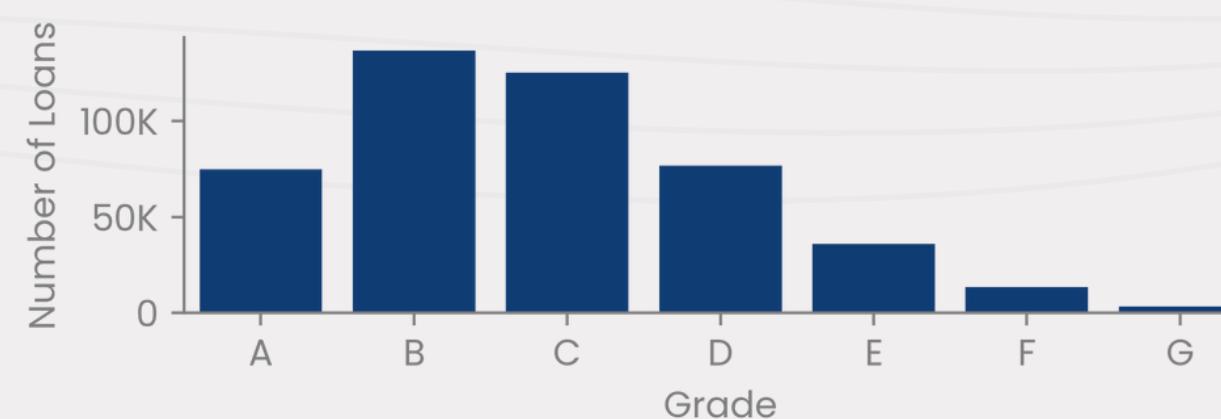
Total Portfolio  
**\$6.7B**

Annual Income  
**\$63K**  
Range: \$2k-\$7,500k

Loan Amount  
**\$12K**  
Range: \$0.5k-\$35k

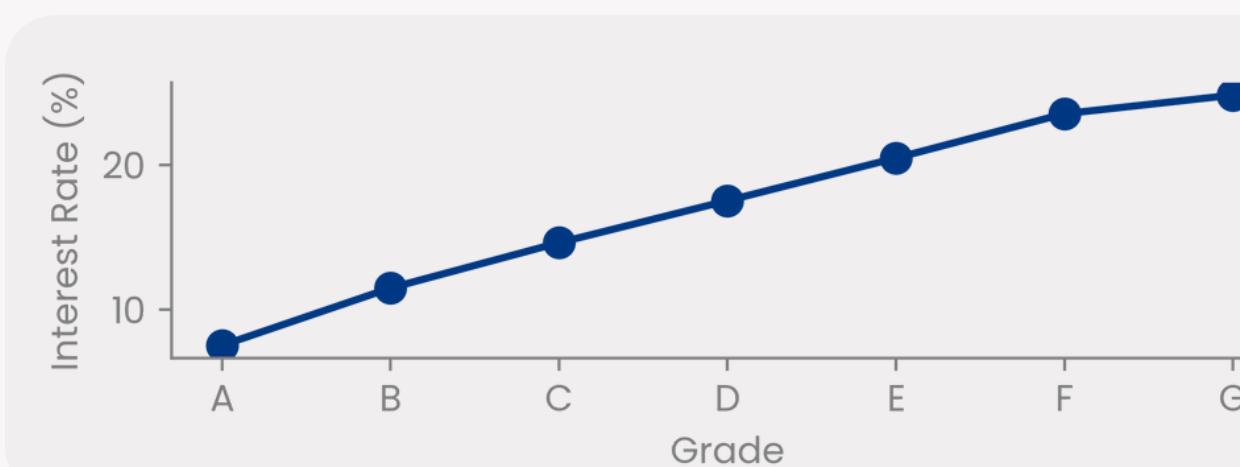
Employment Length  
**10+ years**  
Range: <1 year to 10+ years

# Portfolio Skewed Toward Mid-Grade Risk



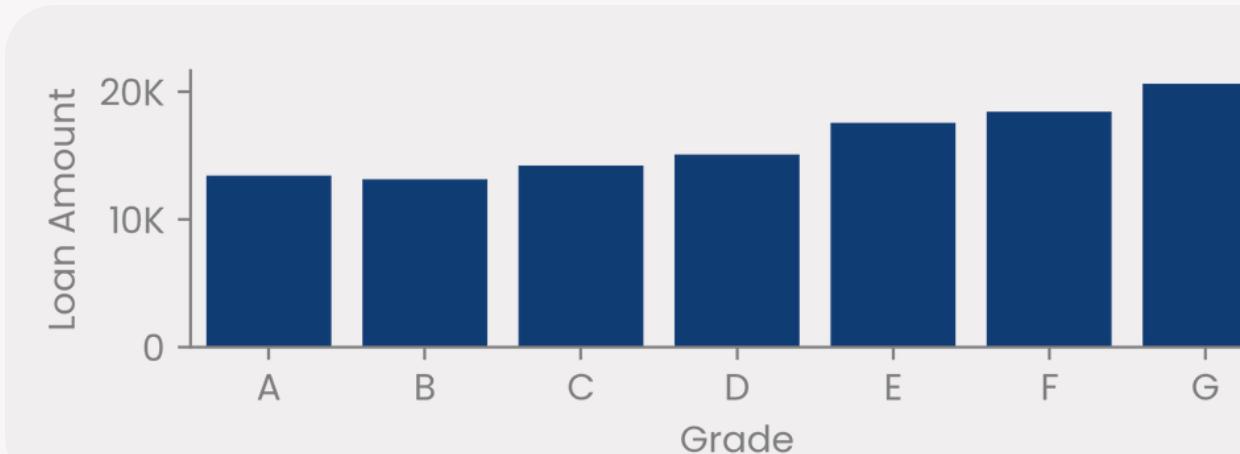
## Heavy Mid-Grade Concentration

Grades B-C represent 56% of portfolio (262K borrowers), while Grade A (lowest risk) only accounts for 16%. This creates significant exposure to medium-risk segments.



## Interest Rate Inefficiency

While rates correlate with risk (7.5% for A to 24.8% for G), the spread between grades isn't proportional—particularly compressed in high-risk grades E-G.



## Inverse Risk-Amount Relationship

Lower grades paradoxically borrow larger amounts—Grade G averages \$20.6K vs Grade A's \$13.4K, amplifying potential losses.

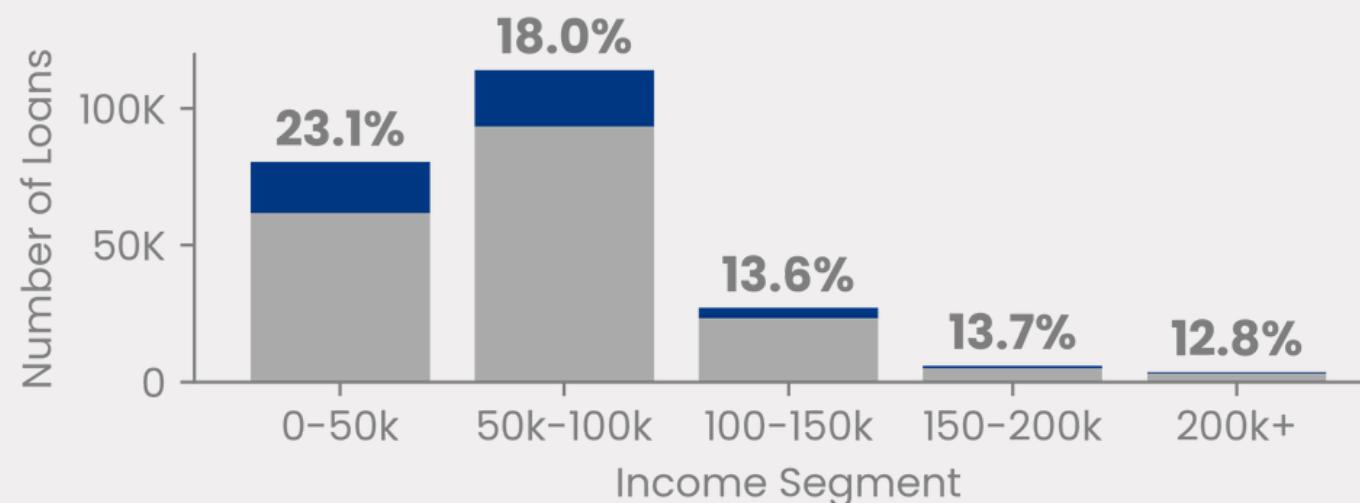
Actionable Item

Rebalance the portfolio by launching targeted campaigns to increase Grade A-B acquisition

# Risk Indicators Analysis: The Poverty-Grade Paradox

## Low Income Drives Half of All Defaults

The 0-50k income segment shows **2x higher default rate** (23.15%) compared to 200k+ segment (12.81%), yet comprises 43% of the portfolio (80,429 loans). This single segment drives **48% of all defaults** (18,617 cases).

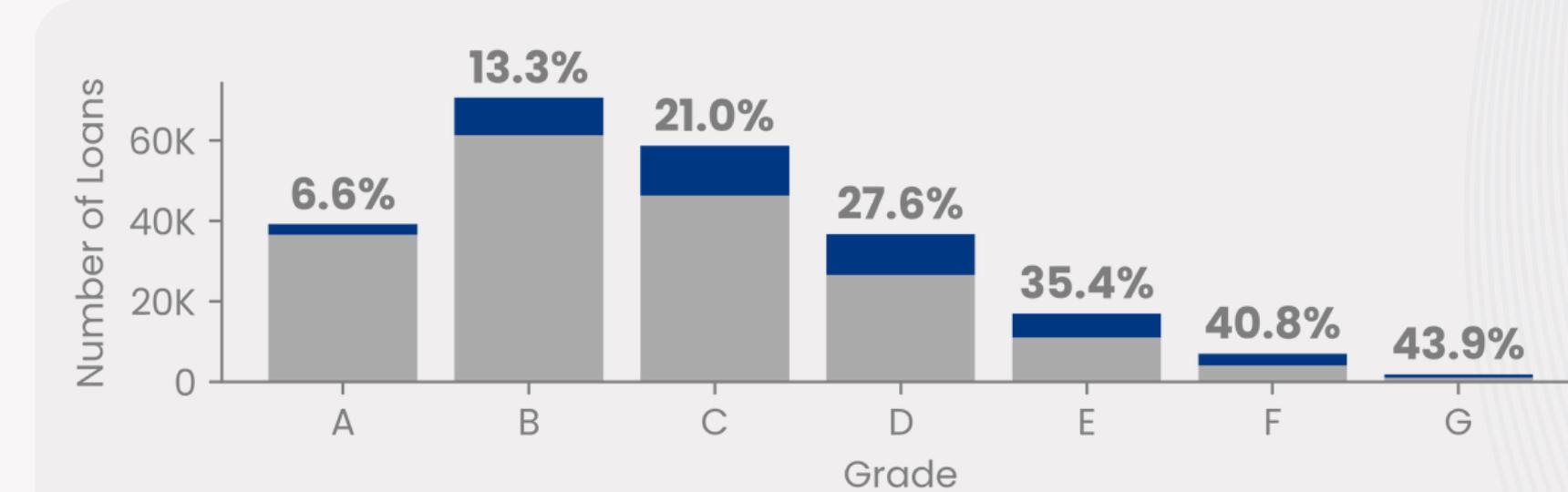


Tighten approval criteria for 0-50k segment

Actionable Item

## Grade D-G: 20% Volume, 50% Defaults

Default rate escalates from 6.6% (Grade A) to 43.9% (Grade G), a **7x gap**. Grades D-G represent only 20% of loans but account for **50% of all defaults** (19,785 cases). Grade D marks the critical inflection point where default rate **jumps to 28%**.



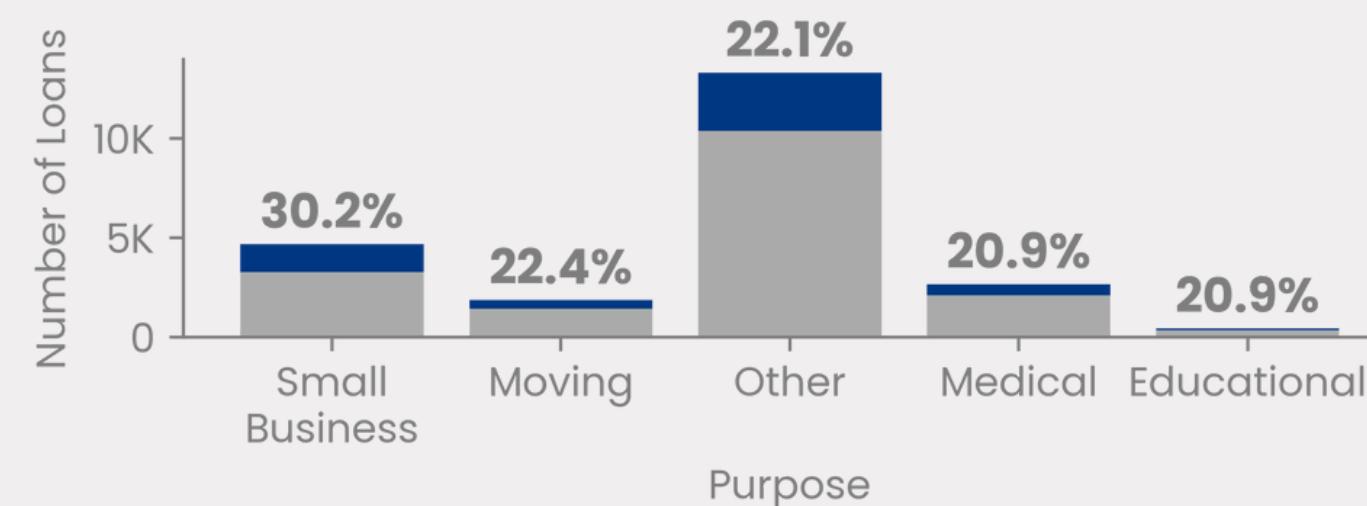
Restrict or discontinue lending to Grades F-G

Actionable Item

# Risk Indicators Analysis: Purpose & Place Risk Nexus

## Small Business Loans: Highest Default Risk

**Small business loans** have 30% default rate, significantly **higher** than other purposes (21-22%). Despite being only 19% of this portfolio subset, they contribute 28% of defaults (1,407 cases). Educational and medical loans show the lowest risk (~21%).

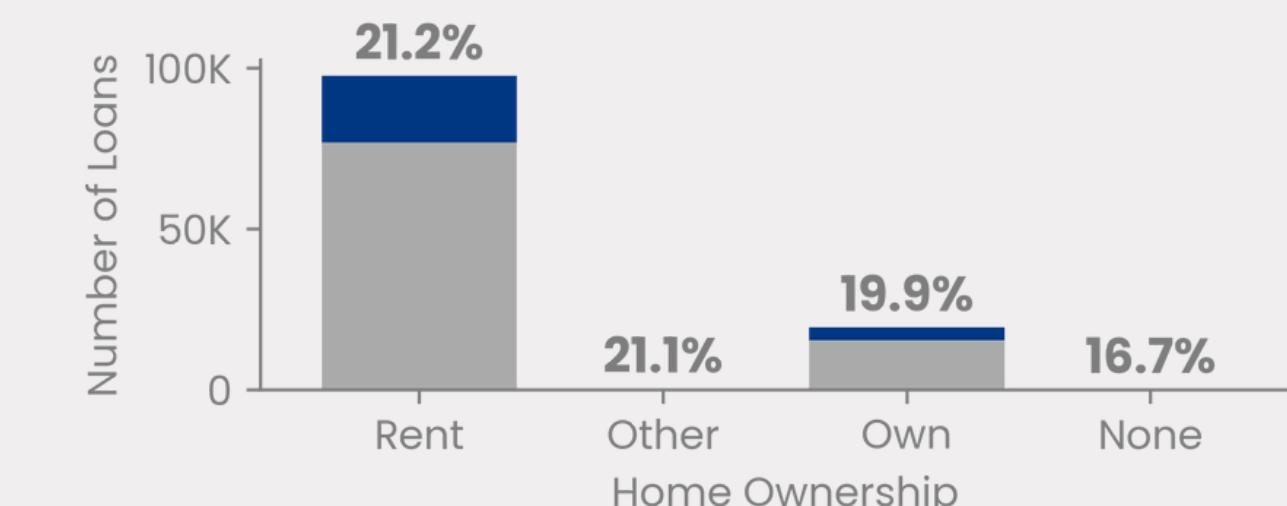


Require cash flow proof for small business loans

 Actionable Item

## Renters Show Highest Default Volume

Default rates are consistent across housing types (20-21%), but **renters dominate 53% of portfolio** and contribute **53% of total defaults** (20,670 cases). Homeowners show slightly lower default rate (19.9%) with better stability indicators.



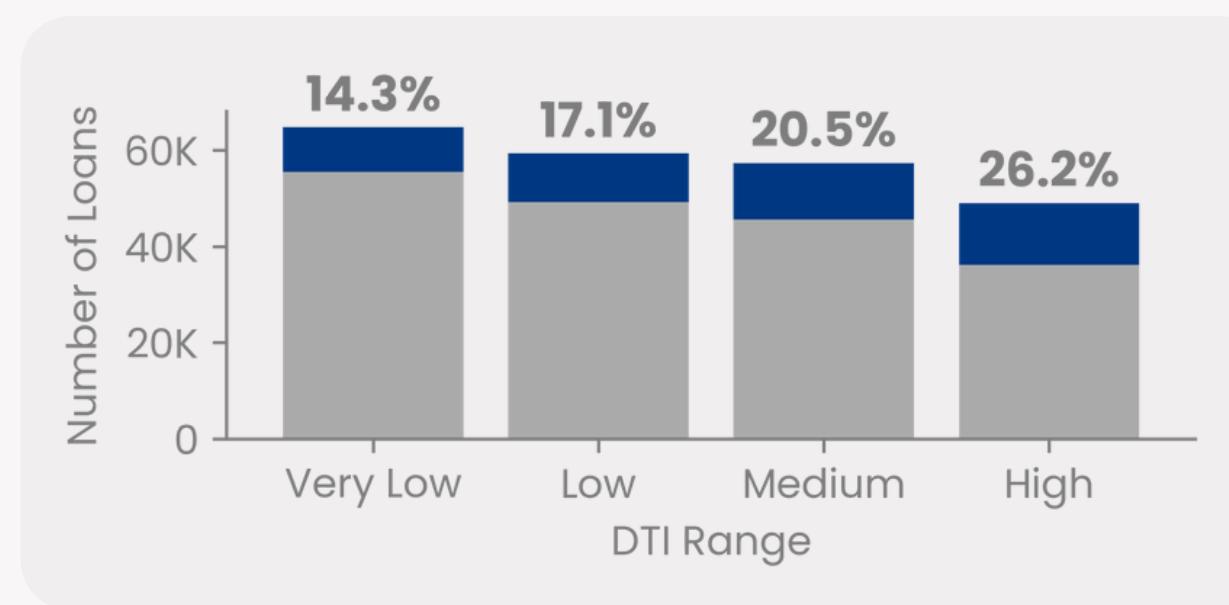
Require higher down payments (15-20%) for renters

 Actionable Item

# Risk Indicators Analysis: The Triple Threat Framework

## High DTI: 26% Default Rate

Default rate climbs from 14% (Very Low DTI) to 26% (High DTI), an **83% increase**. High DTI borrowers are 25% of portfolio but **drive 29% of defaults** (12,869 cases).

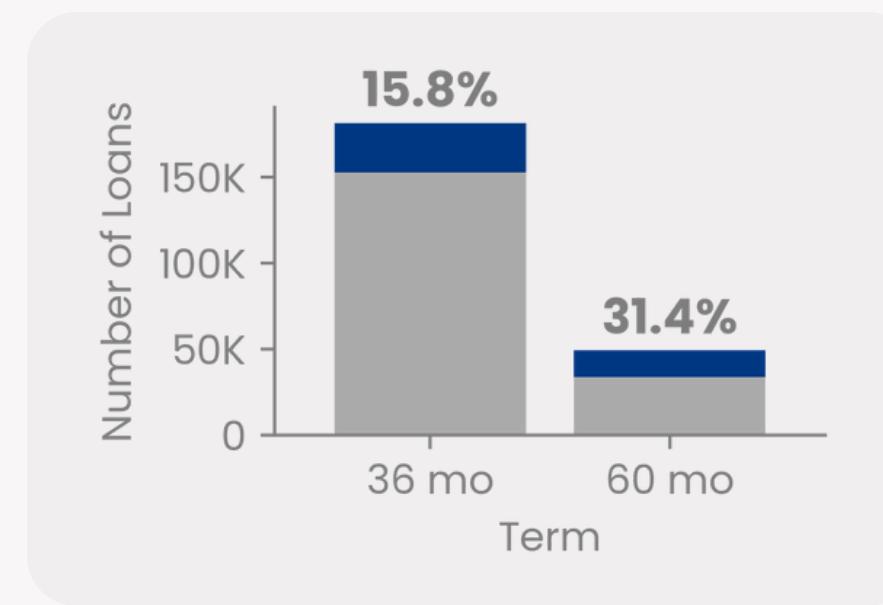


Cap DTI at 35-40% maximum

Actionable Item

## 60mo: 2x Default Risk

60-month loans default at 31% vs 16% for 36-month, a **2x gap**. They're 25% of volume but **35% of defaults**.

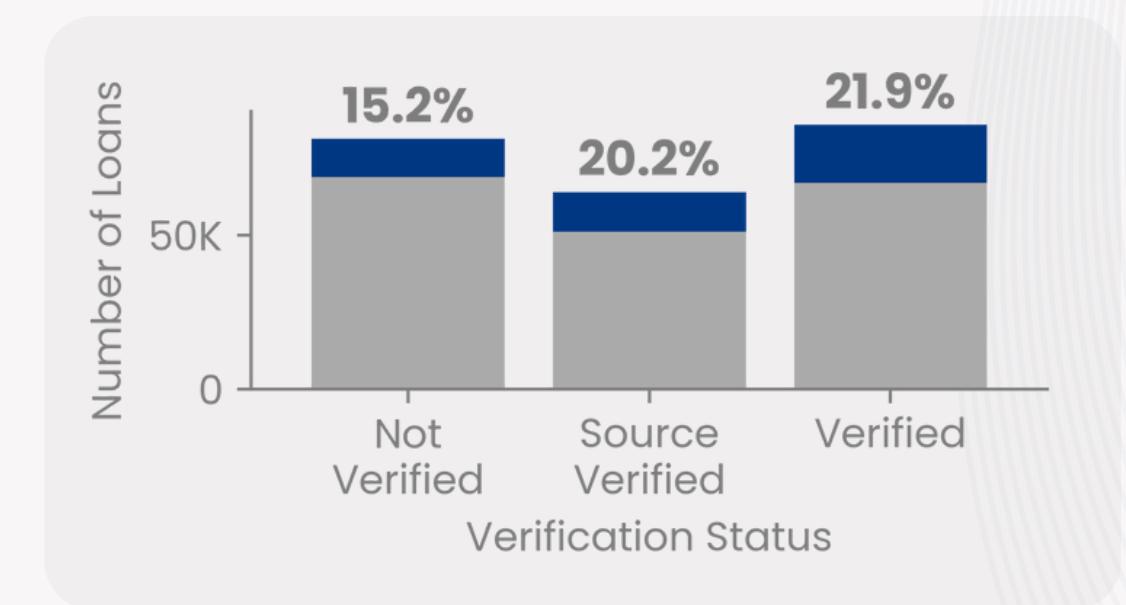


Limit 60mo terms to Grade A-B

Actionable Item

## Verified Loans Default More

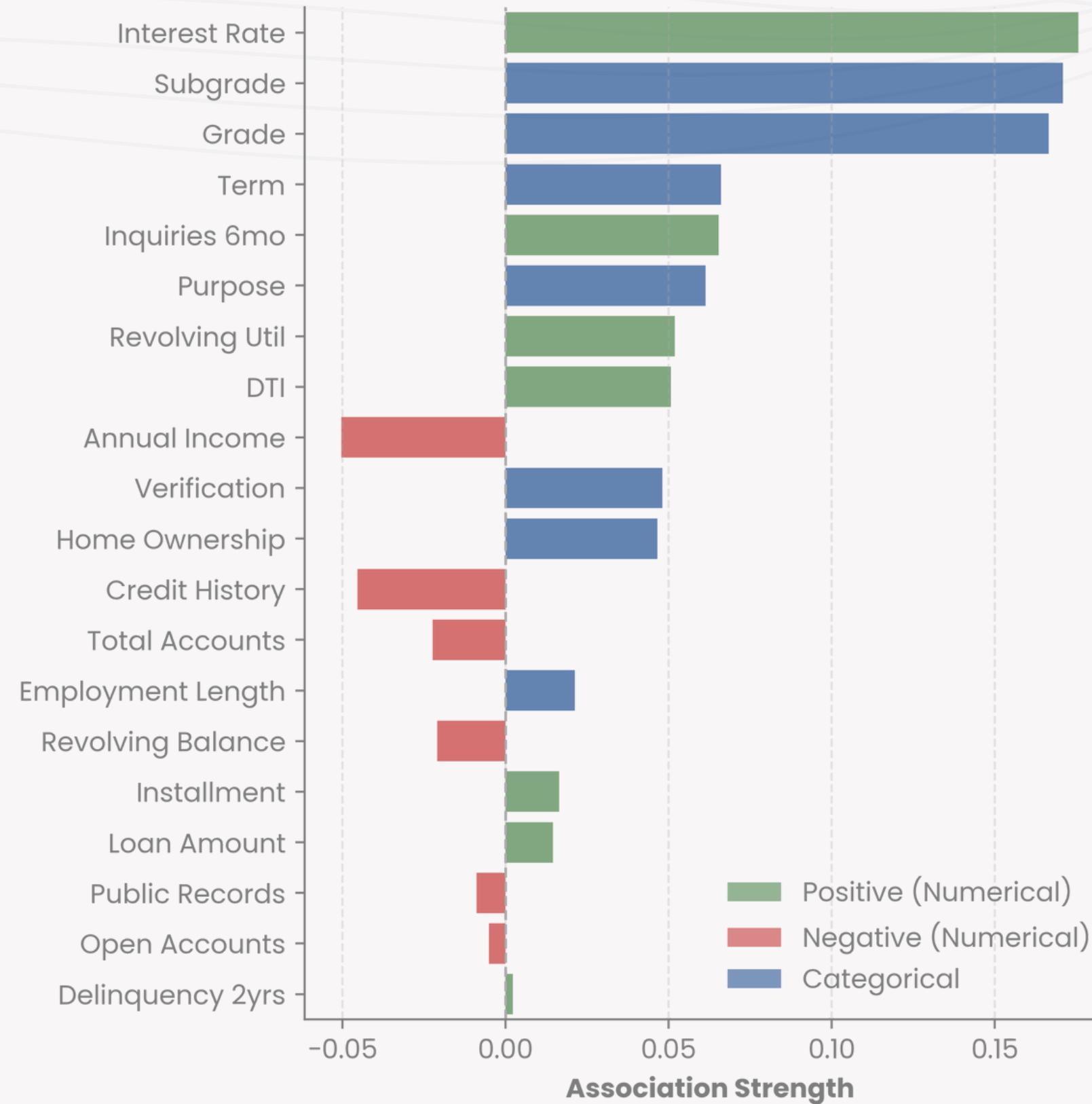
**Verified loans** show **22% default rate** vs 15% for unverified, a counterintuitive 44% increase. Verified loans are 44% of portfolio but contribute 43% of defaults.



Don't rely on verification alone

Actionable Item

# Feature Correlation Analysis



## Weak Linear Relationships Across All Features

 Interest rate (0.176), sub\_grade (0.171), and grade (0.167) are the top correlated features with default risk. However, these values below 0.20 are classified as weak correlations, indicating limited linear predictive power when examined individually.

## Negligible Individual Predictors

 The remaining features show correlations below 0.10—term (0.066), recent inquiries (0.065), loan purpose (0.061), DTI (0.051), and annual income (-0.050). These fall into the "very weak" category, demonstrating minimal standalone impact on default prediction.

### Key Insight

Correlations below 0.3 indicate very limited linear relationships. Credit default appears to be driven by complex interactions between multiple factors rather than any single dominant predictor. This suggests multivariate analysis and advanced modeling techniques will be necessary.

# Data Preprocessing Pipeline

## Numerical Features

Iterative Imputer, Winsorize, Yeo-Johnson Transform



## Ordinal Features

Custom Ordinal Encoder (grade, sub\_grade, etc)



## Categorical Features

Label Encoding (2 cols), OneHot Encoding (7 cols)



## Outlier Handling

Winsorization at 1st and 99th percentiles to clip extreme values while preserving distribution



## Missing Data

IterativeImputer for numerical (MICE algorithm) + SimpleImputer for categorical (most frequent)



## Distribution Normalization

Yeo-Johnson transformation handles skewed distributions and works with zero/negative values



## Custom Ordinal Encoding

Hierarchical mapping for grade (A-G), sub\_grade (A1-G5), employment length (0-10+ years)

## Feature Selection

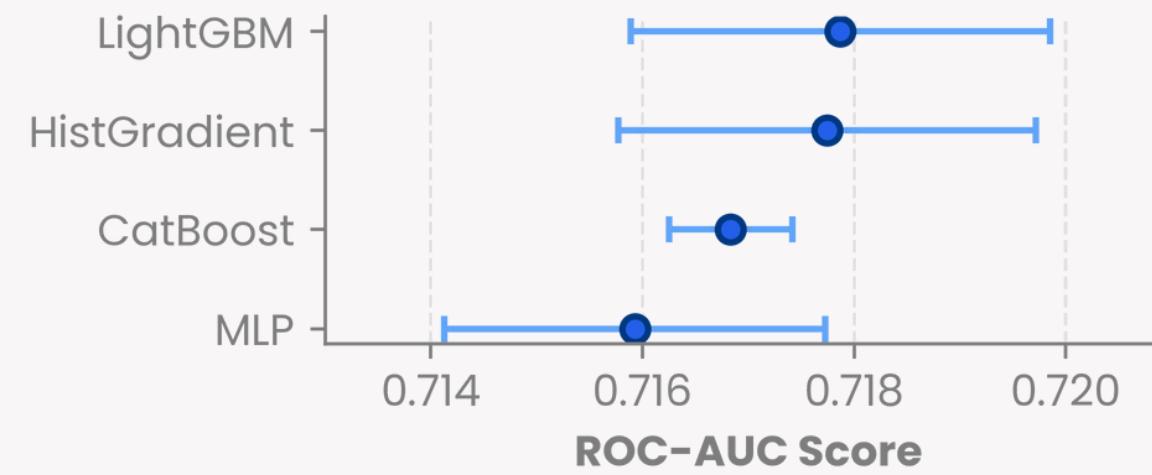
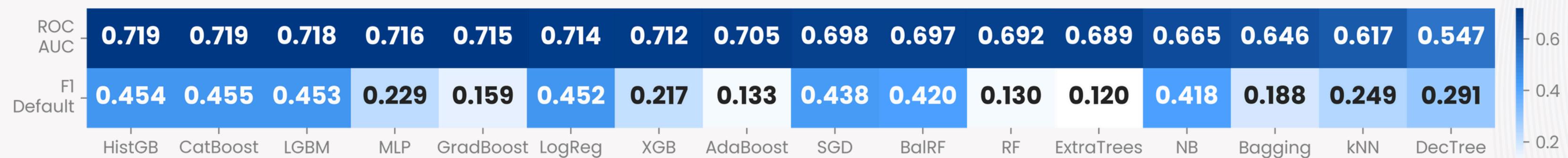
- L1 Regularization (Lasso) for linear models
- Feature Importance (Mean) for tree ensemble
- XGBoost Importance (Median) for boosting
- Variance Threshold (0.01) for distance-based



# Model Selection

## Initial Model Screening on Validation Set

Initial evaluation of **16 machine learning algorithms** showed ROC-AUC performance ranging from **0.547 to 0.719**, with **HistGB**, **CatBoost**, **LGBM**, and **MLP** as top performers. The four best models were then selected for **more comprehensive cross-validation** to validate performance stability. This approach optimizes computational efficiency while maintaining evaluation quality.

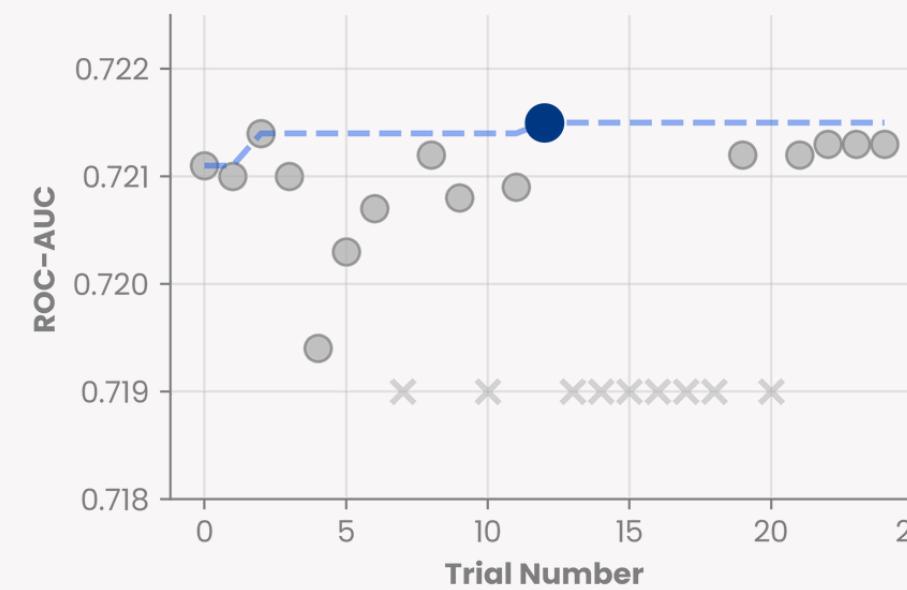


## Cross-Validation Results (5-Fold with 95% CI)

**LightGBM** was selected as the final model with a CV ROC-AUC of **0.7179** (highest) and a **stable confidence interval** (0.7159 – 0.7199). This model provides an **optimal balance** between **prediction performance** and **cross-fold consistency**, with **good recall** for positive case detection.

# Model Optimization

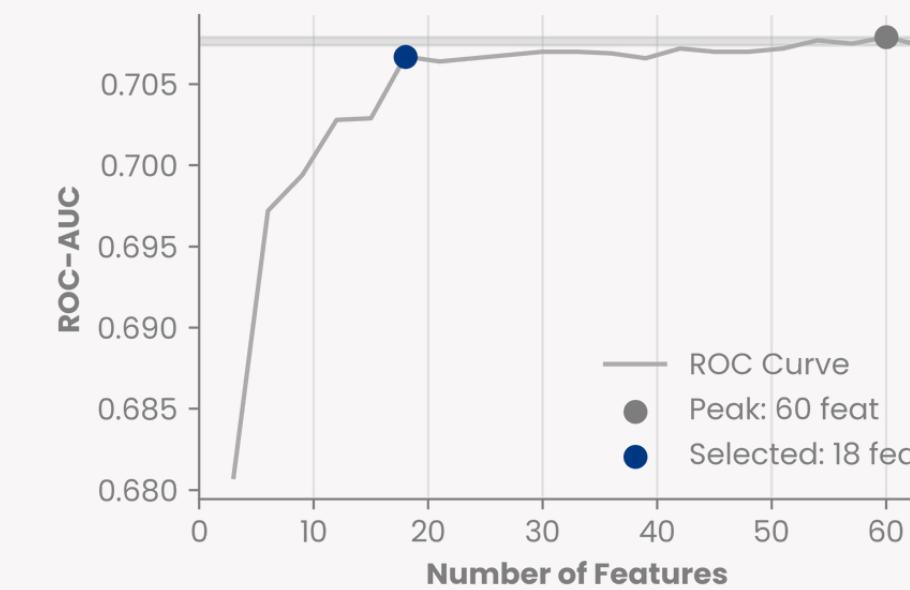
## Hyperparameter Tuning



**ROC-AUC: 0.7215**

The optimization process completed **16 trials** and **pruned 9 others** over 5.2 hours using the **TPE algorithm**, achieving a best ROC-AUC score of **0.7215**.

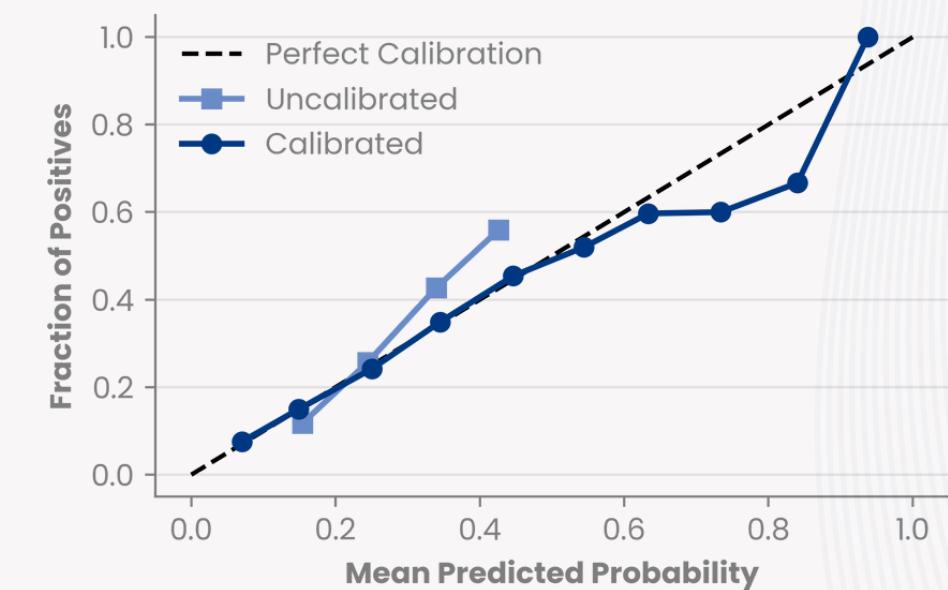
## Feature Optimization



**-70% Reduction**

Feature reduction from **60 to 18 features** (70% reduction) **maintained 99.83% performance**, with a final ROC-AUC of 0.7067 and training time of 6.2 seconds.

## Model Calibration



**Isotonic Regression**

The model uses **Isotonic Regression** with 5-Fold Cross Validation, improving **Brier Score** and **Log Loss** for reliable probability predictions in production.

# Evaluation

## ROC-AUC

**0.708**

Acceptable

## Brier Score

**0.152**

Excellent

## Log Loss

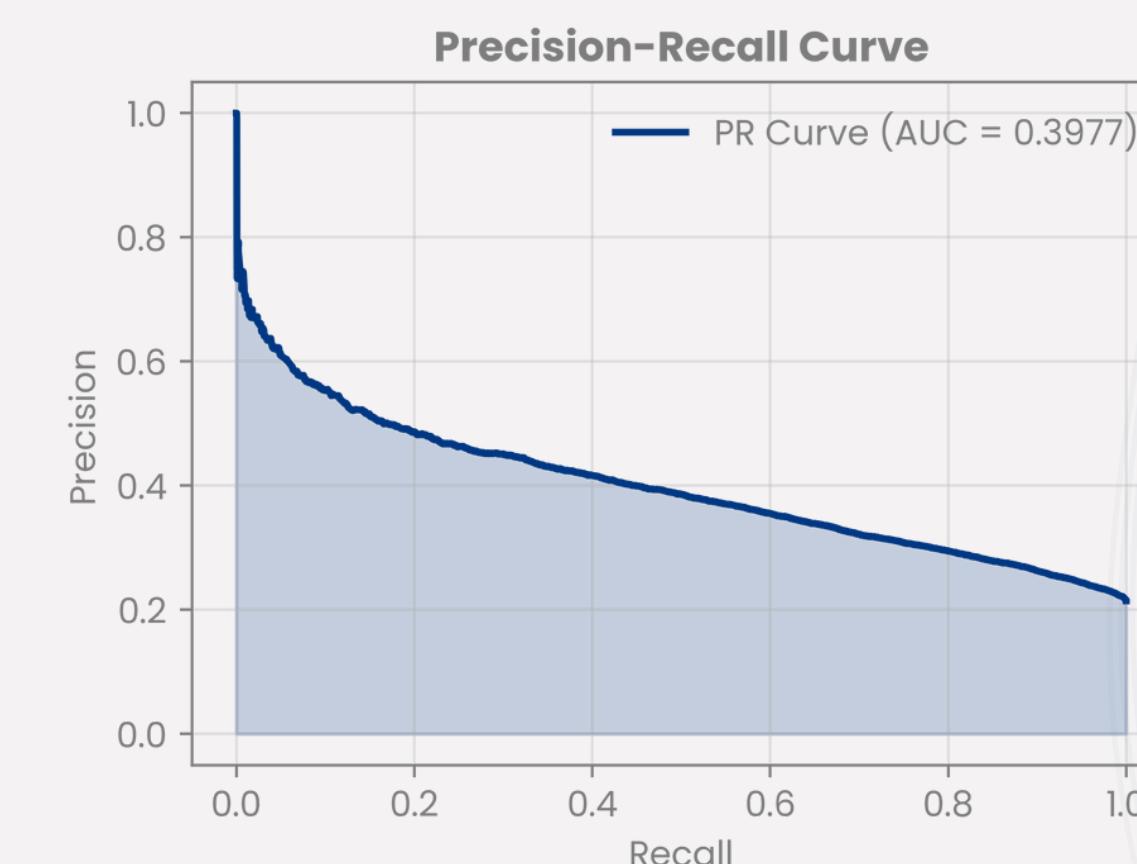
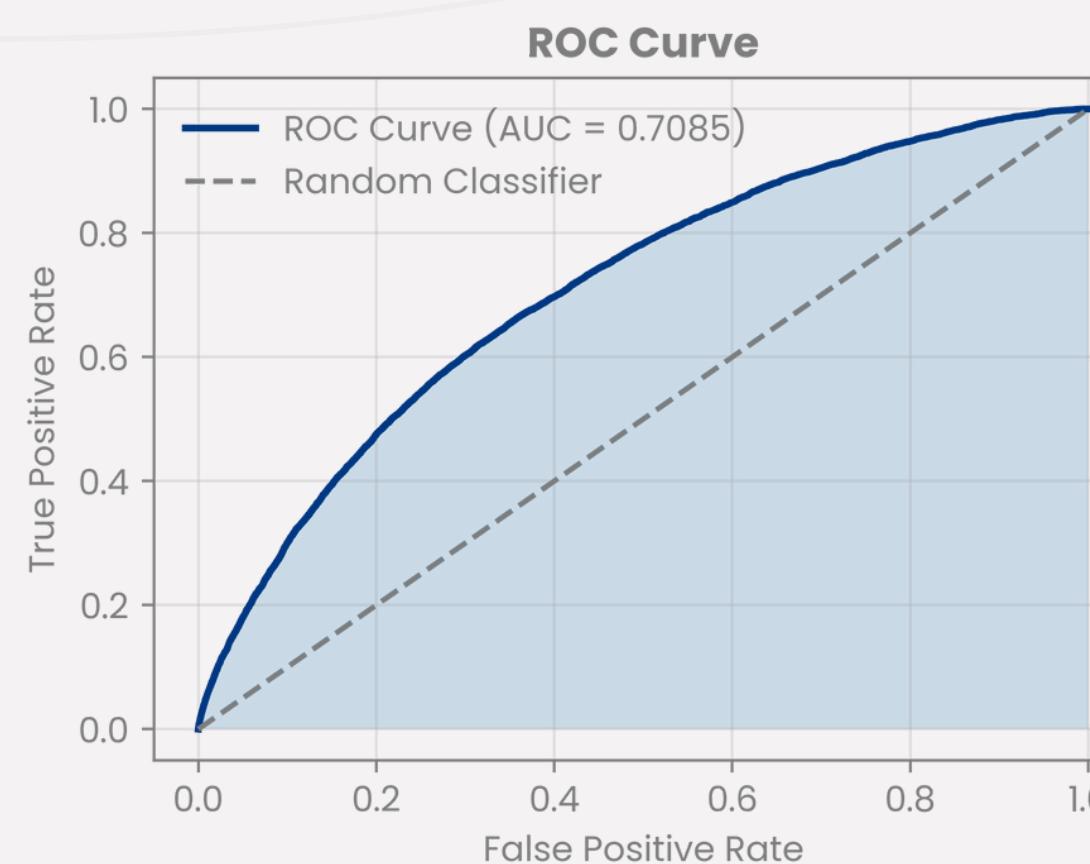
**0.472**

Fair

## Accuracy

**78.98%**

Good



The model demonstrates **solid discriminative ability** with ROC-AUC of 0.708 and **well-calibrated probabilities** (Brier Score: 0.152). Overall accuracy reaches 78.98%. The lower PR-AUC (0.3977) reflects class imbalance, addressed through balanced class weights. **Tier segmentation** is applied for practical implementation.

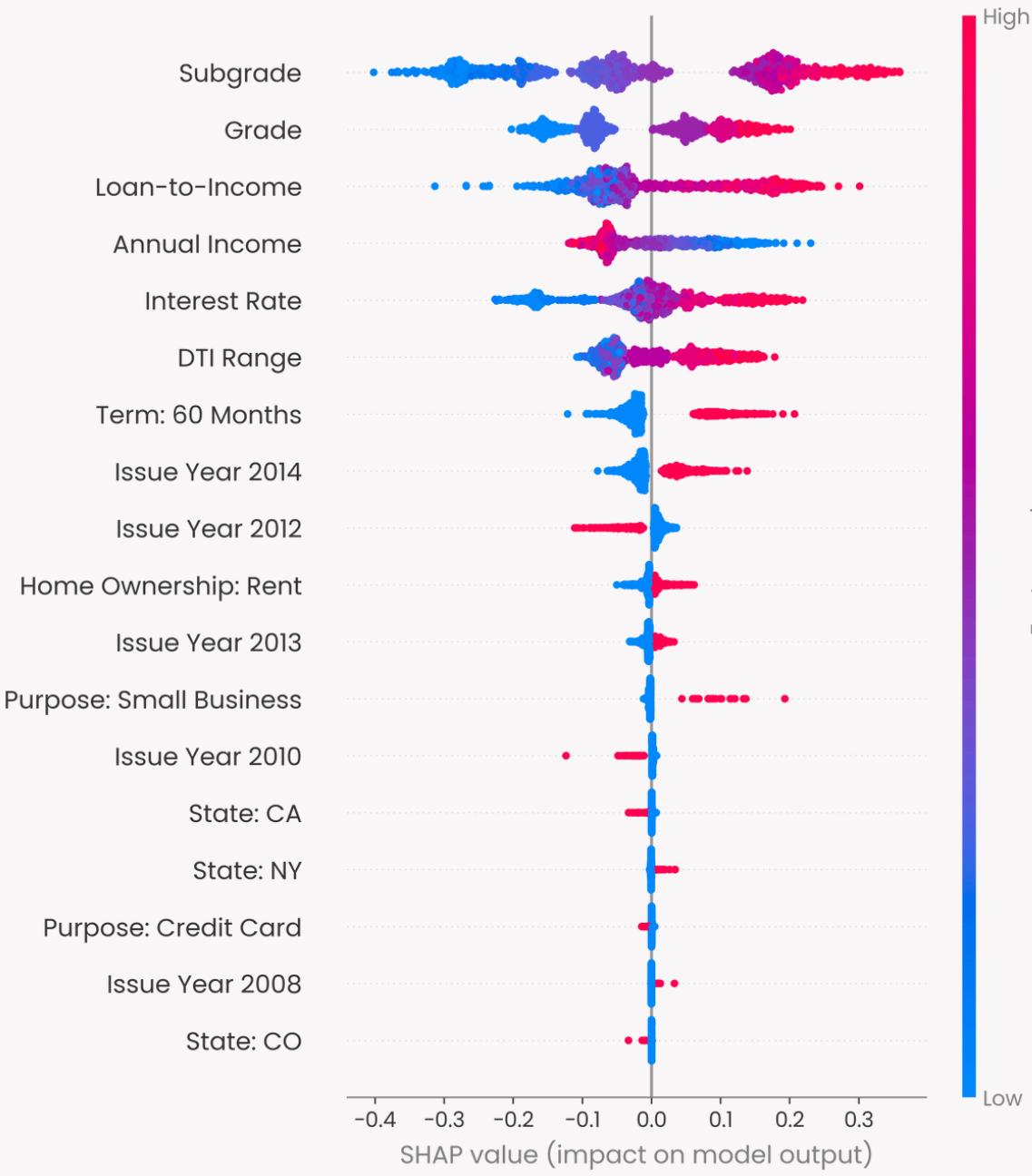
# Model Interpretability Overview

## Why SHAP Analysis?

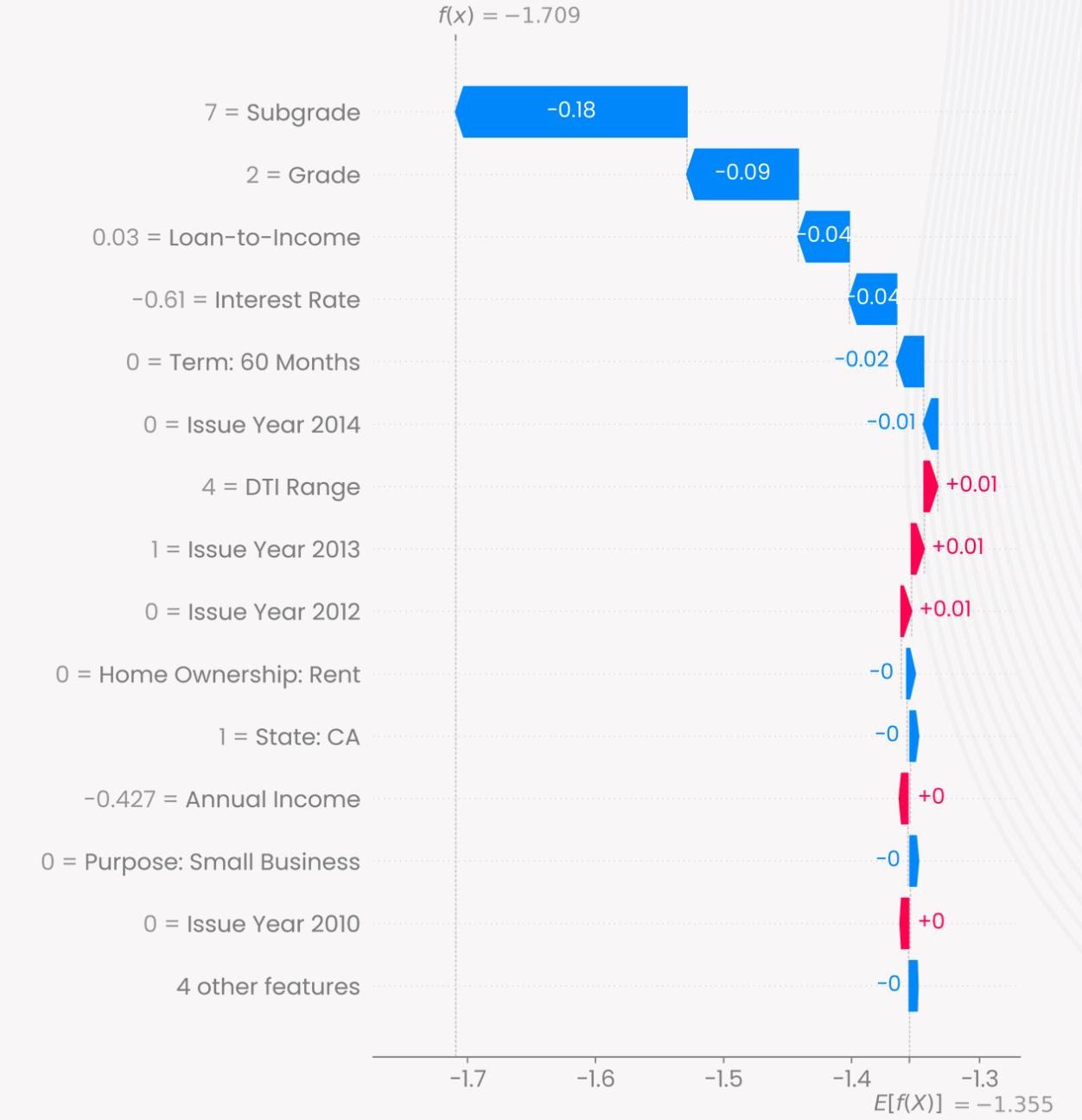
- \* Regulatory compliance (OJK transparency requirements)
- \* Transparent explanations for credit officers & applicants
- \* Validate model learns correct risk patterns

**DTI Range** and **2012–2014 vintage loans** show highest risk contribution, while **premium grades** (A–B) are strongest protective factors against default.

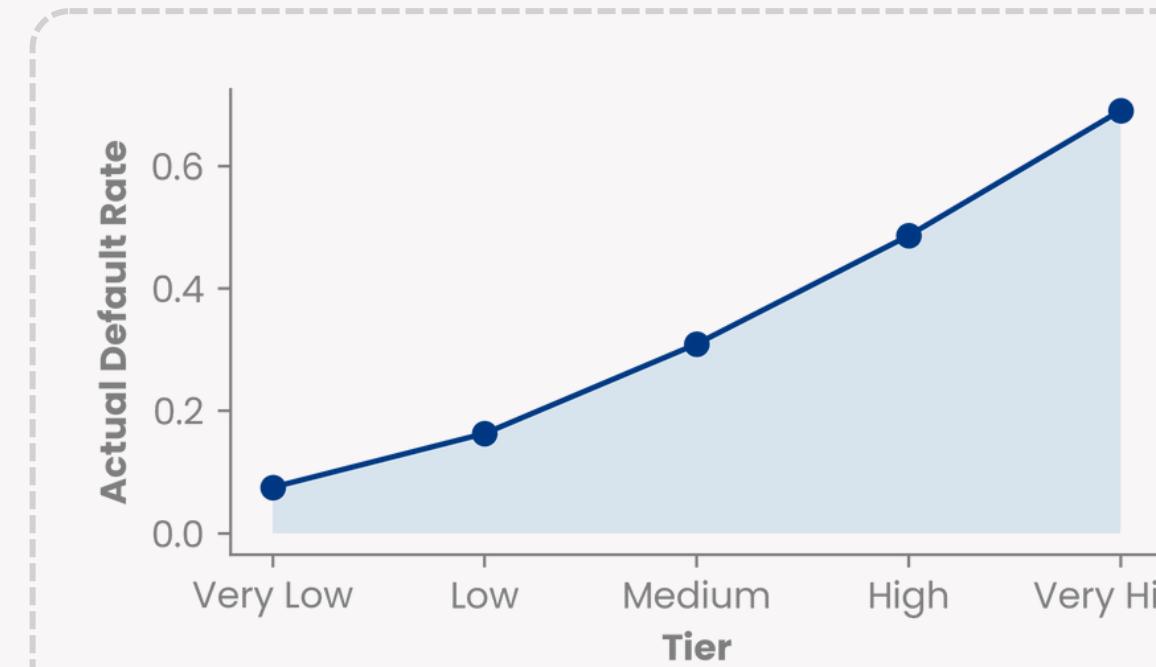
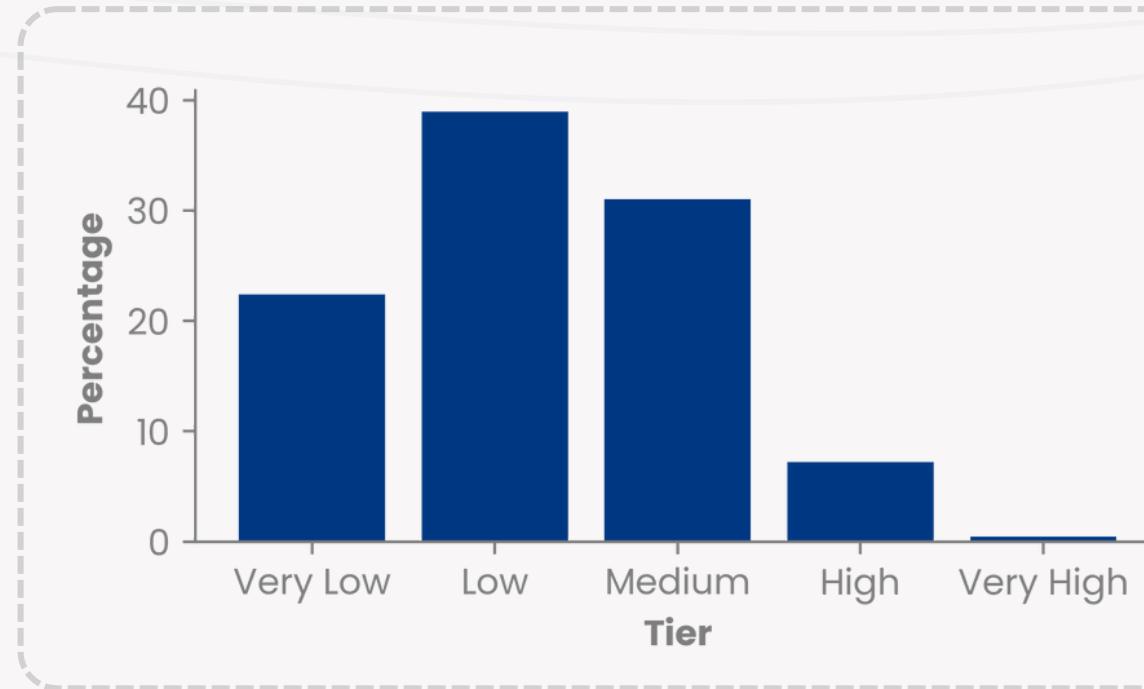
## Global Feature Importance



## Individual Case Example



# Risk Tier Segmentation



## Risk Tier Actions

 <b>Very Low</b>	Auto-approve	 <b>Low</b>	Standard Approval	 <b>Medium</b>	Manual Review
 <b>High</b>	Strict Review	 <b>Very High</b>	Reject		

	Threshold	Volume	Default Rate	Predicted Prob
<b>Very Low</b>	11%	10,644	7.5%	7.4%
<b>Low</b>	23%	18,514	16.3%	16.3%
<b>Medium</b>	43%	14,751	30.9%	30.9%
<b>High</b>	62%	3,423	48.6%	49.6%
<b>Very High</b>	100%	207	69.1%	66.5%



Five-tier segmentation achieves strong calibration with <2% deviation, enabling automation of 20% approvals and focused manual review on 10% high-risk cases.

# Financial Impact

**\$16.0M**

Portfolio Loss Reduction  
(31.5% decrease)

**19.6%**

Default Rate  
(8.8pp reduction)

**2.22%**

Risk-Adjusted Return  
(+95 basis points)

# Executive Summary

The developed credit risk predictive model achieved solid performance with **ROC-AUC score of 0.708** and **Brier Score of 0.152**. Implementation of the **five-tier risk segmentation system** enables automated decisions for **20% of low-risk applications** and focused manual review on **10% of high-risk cases**.

Portfolio analysis revealed critical findings: the **0-50k income segment contributes 48% of total defaults** despite being only 43% of the portfolio, while **Grades D-G** representing just 20% of volume **account for 50% of all defaults**. Small business loans show the **highest default rate (30%)**, and **60-month term loans carry 2x the risk** compared to 36-month terms.

The model delivers significant financial impact with potential **portfolio loss reduction of \$16.0M (31.5% decrease)**, **default rate reduction to 19.6%** (down 8.8%), and **risk-adjusted return improvement to 2.22%** (up 0.95%). Strategic recommendations include tightening approval criteria for low-income segments, restricting lending to Grades F-G, and implementing a tier-based decision framework for portfolio risk optimization.

# Thank You

This presentation is the result of a Project-Based Virtual Internship as  
Data Scientist at ID/X Partners x Rakamin Academy



View Project on GitHub

---

E-mail

[afsilmis@gmail.com](mailto:afsilmis@gmail.com)

Connect

[linkedin.com/in/az-zukhrufu-fi-silmi-suwondo/](https://linkedin.com/in/az-zukhrufu-fi-silmi-suwondo/)

Portfolio

[github.com/afsilmis/](https://github.com/afsilmis/)