

# 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

**466K**

records

**75**

features

**780**

MB

2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014

## Target Distribution



### Good Loans

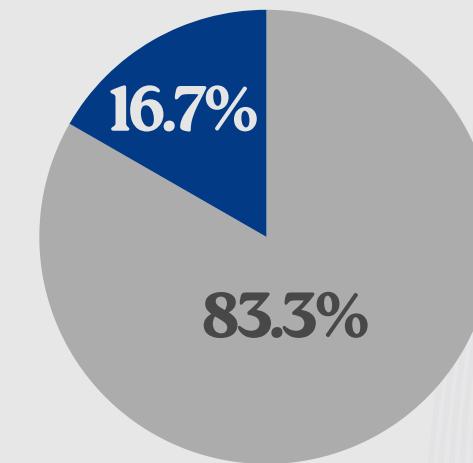
(Fully Paid, Current, In Grade Period)



### Bad Loans

(Charged Off, Default, Late)

\*Includes policy-exception loans classified by final outcome



## Feature Categories

### Identification & Admin



5 features

### Loan & Borrower Info



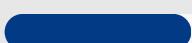
17 features

### Credit & Payment History



40 features

### Dates & Timeline



5 features

### Status & Target



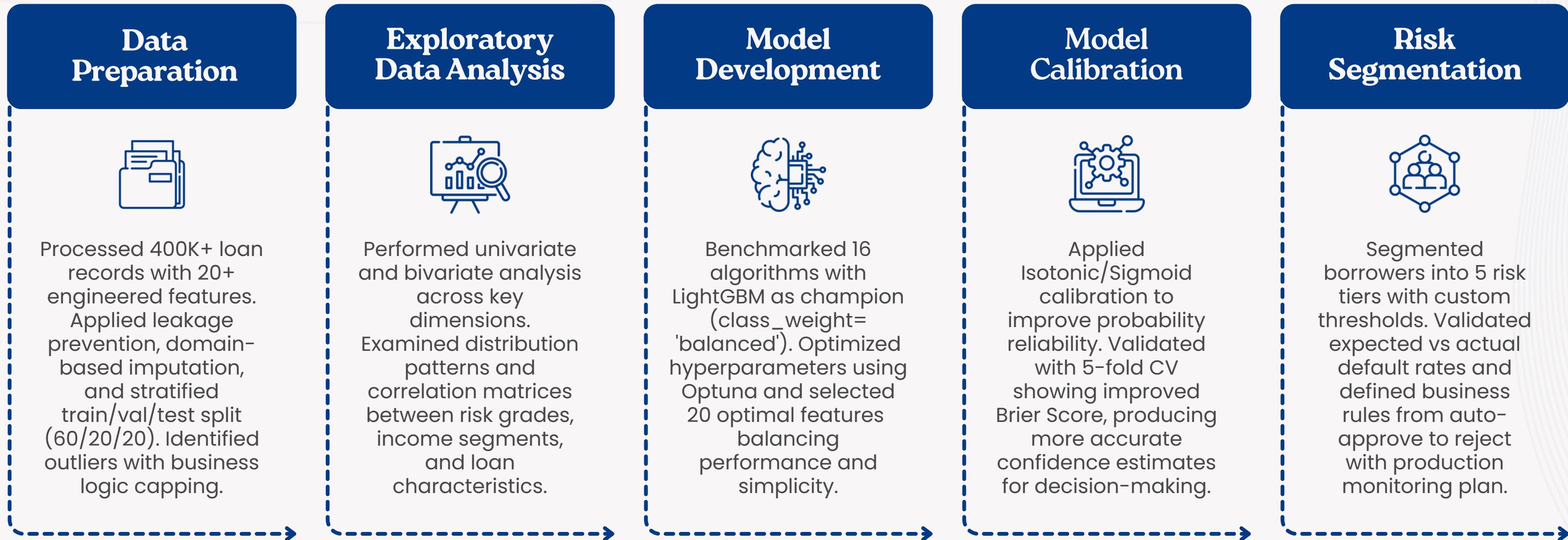
4 features

### Joint Application



3 features

# 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

 Total Loans

**466K**



 Total Value

**\$6.7B**



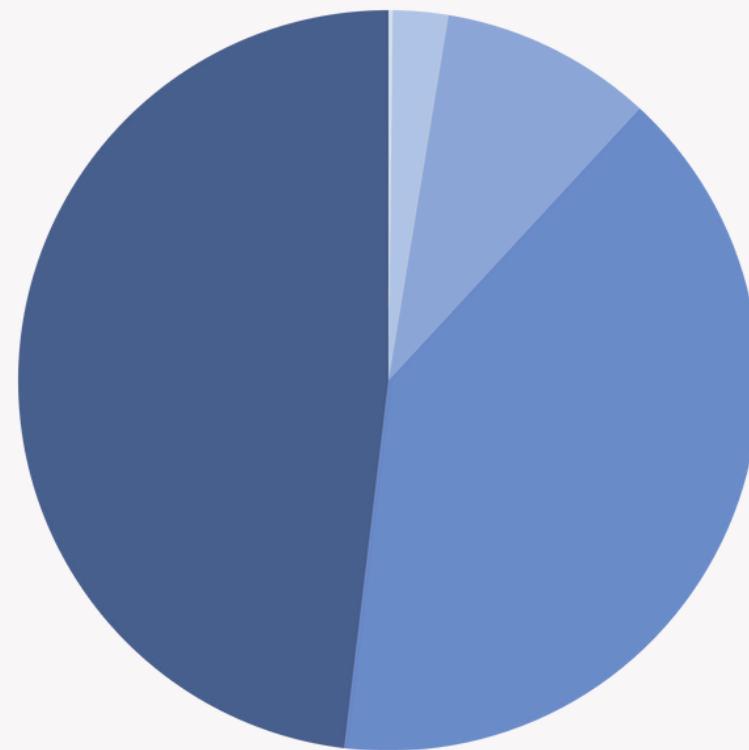
 Average Loan

**\$14K**



 Healthy Portfolio

**87.7%**



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

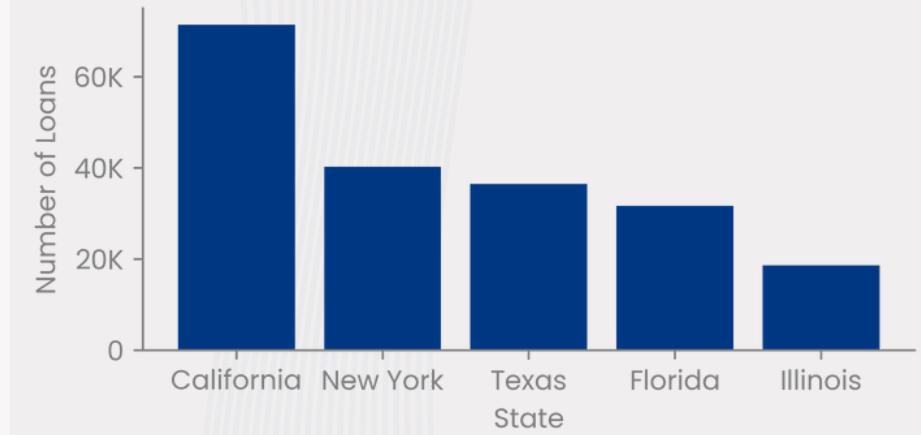
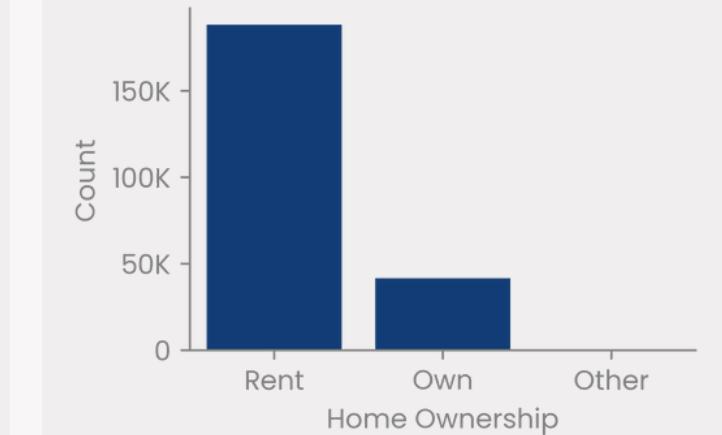
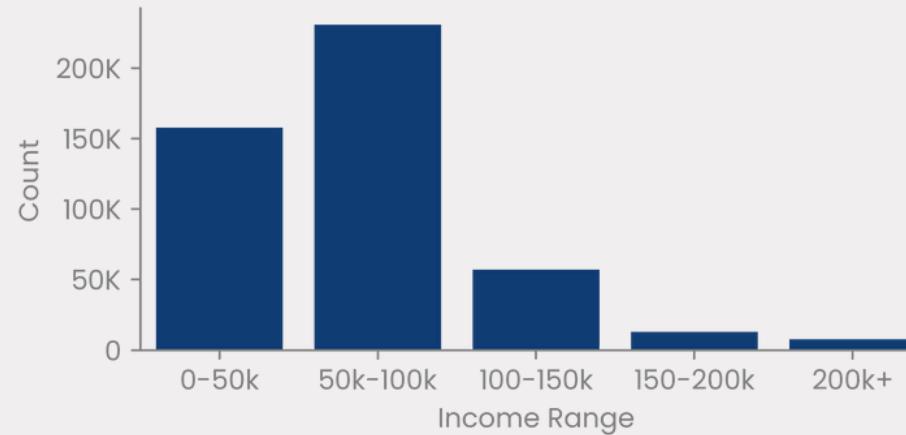
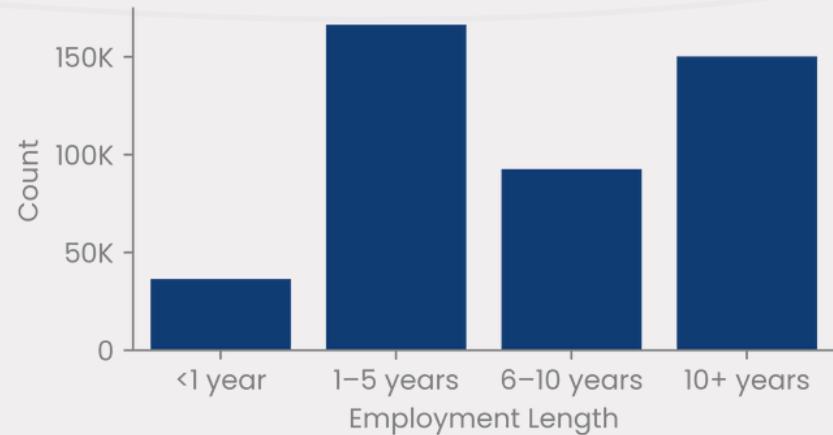
 Actionable Item

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**

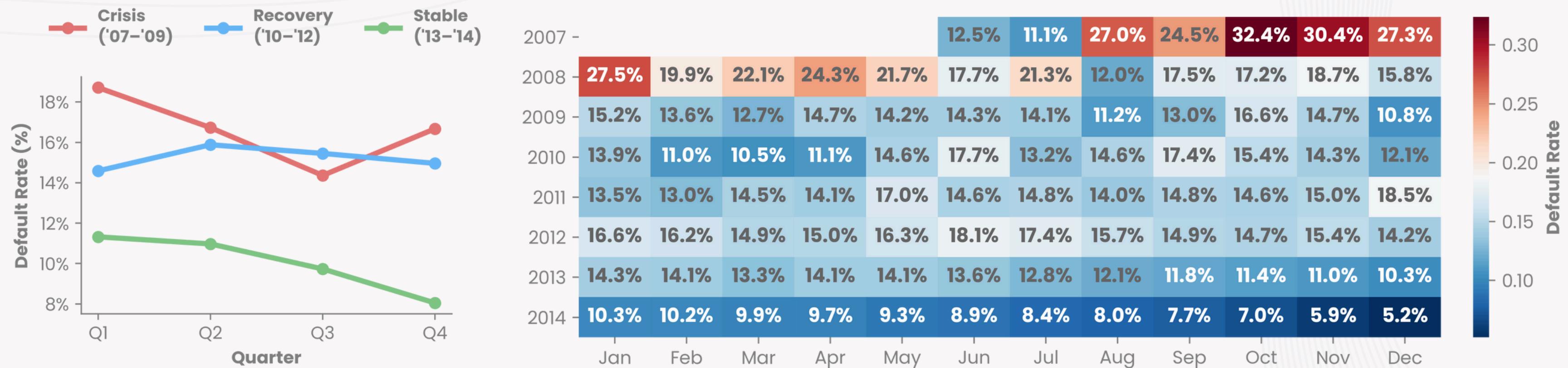
Total Portfolio  
 **\$6.7B**

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

Loan Amount  
 **\$14K**  
Range: \$0.5k-\$35k

Employment Length  
 **1-5 years**  
Range: <1 year to 10+ years

# Seasonality Analysis



Monthly default rates were grouped into three regimes based on risk levels. Crisis regime exhibits the highest default rates (14-19%), Recovery shows moderate levels (15-16%), and Stable represents the lowest risk period (8-11%). This separation allows us to distinguish between macroeconomic conditions and temporal patterns.

 Strong seasonal pattern exists across all regimes, with Q1 consistently showing highest default rates and Q4 the lowest

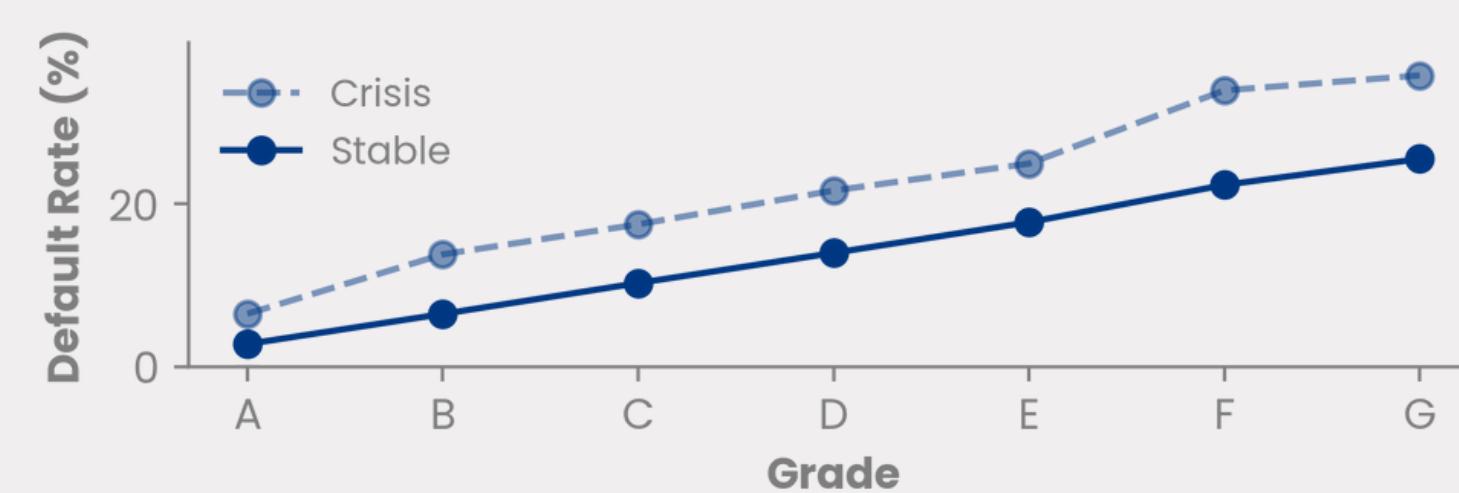
 The seasonal pattern remains consistent regardless of economic regime, indicating independence between macroeconomic and seasonal factors

 The gap between Q1 and Q4 default rates is proportionally similar across all regimes

# Risk Analysis: Credit Grade & DTI Paradoxes

## The Grade Hierarchy Illusion

Default rate recovery is uneven: **premium grades (A-C)** **improved far better** than high-risk grades (F-G). The A-G default gap narrowed from 29 to 23 points, favoring **high-quality borrowers**. **Grades D-E** are the **sweet spot** (14-18% default) for higher yields than A-C.

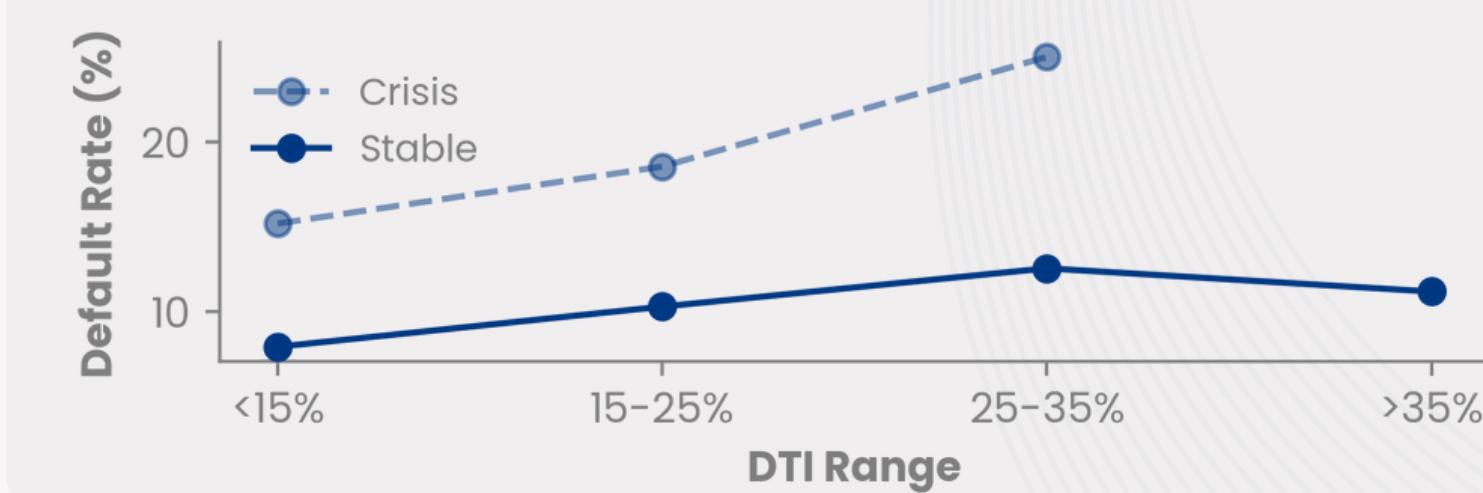


Actionable Item

Shift portfolio focus to D-E. Review pricing or gradually exit the F-G segment as the margin doesn't justify the risk.

## The DTI Sweet Spot Anomaly

**Low DTI (<15%)** is the **safest**. Surprisingly, in the **Stable** period, **DTI >35%** performed better (11.2%) than DTI 25-35% (12.5%). This suggests **income verification** or **high-income borrowers manage high debt better**, contradicting the high-DTI-high-risk assumption.



Actionable Item

Prioritize DTI <15%. For DTI >35%, use extra behavior checks instead of auto-rejecting to target high-income borrowers.

# Risk Analysis: Income & Tenor Dynamics

## Income Risk Reversal Pattern

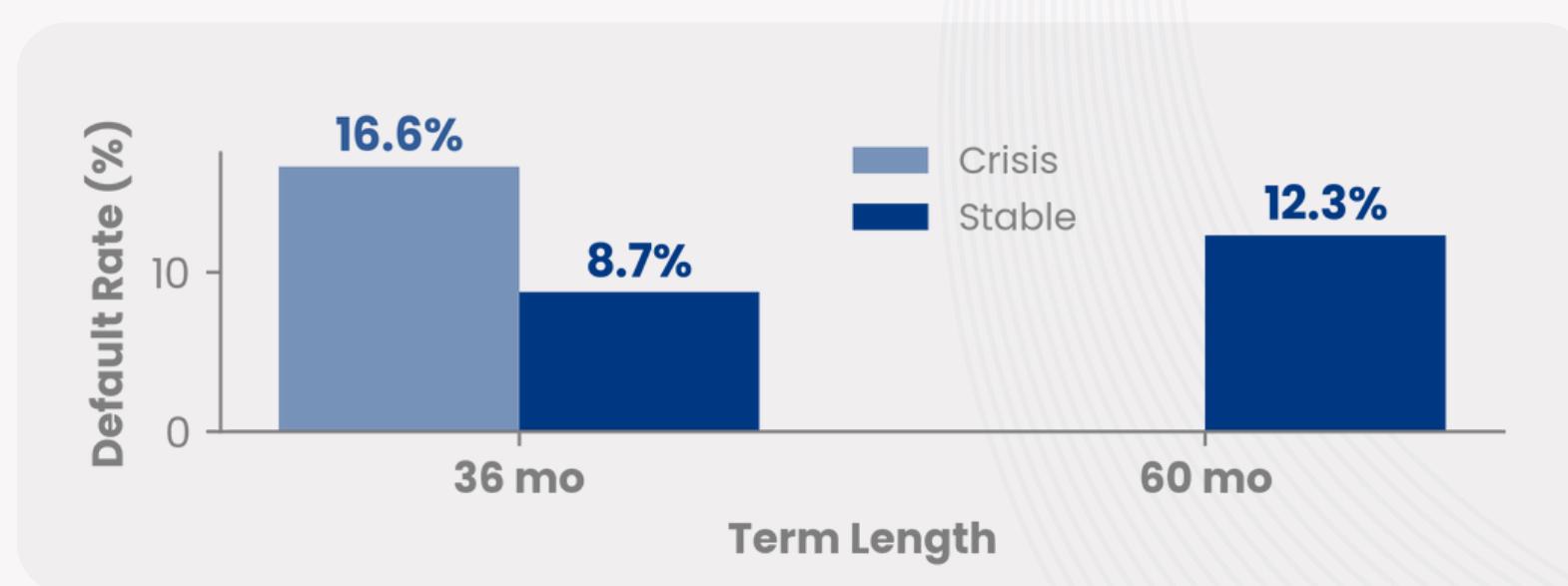
Risk pattern reversed: **High-income (150K+)** was riskier than mid-income (50-150K) during the **Crisis**. In **Stable** times, the **pattern normalized** (higher income = lower default, 12.1% to 6%). The 200K+ segment saw the best 70% improvement.


↗ Actionable Item

Expand aggressively into 100K+. Use stress tests for high-income during downturns (focus on debt).

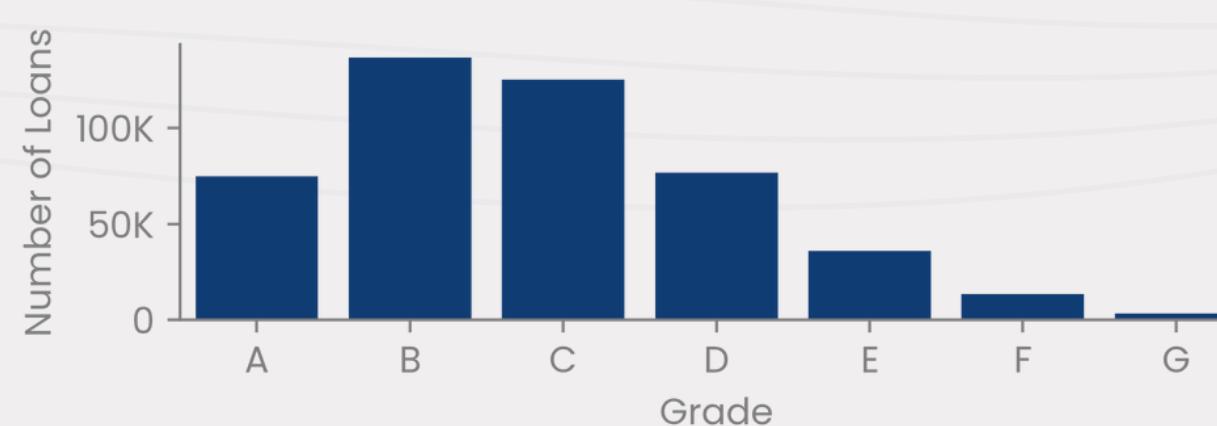
## The 60-Month Premium Penalty

The **60-month term** (only in Stable period) defaulted at 12.3%, 41% **higher** than the 36-month term (8.7%). This confirms **longer tenor equals higher risk**, likely due to extended exposure to life events. The **36-month term** was **more resilient**, dropping 47% from Crisis to Stable.


↗ Actionable Item

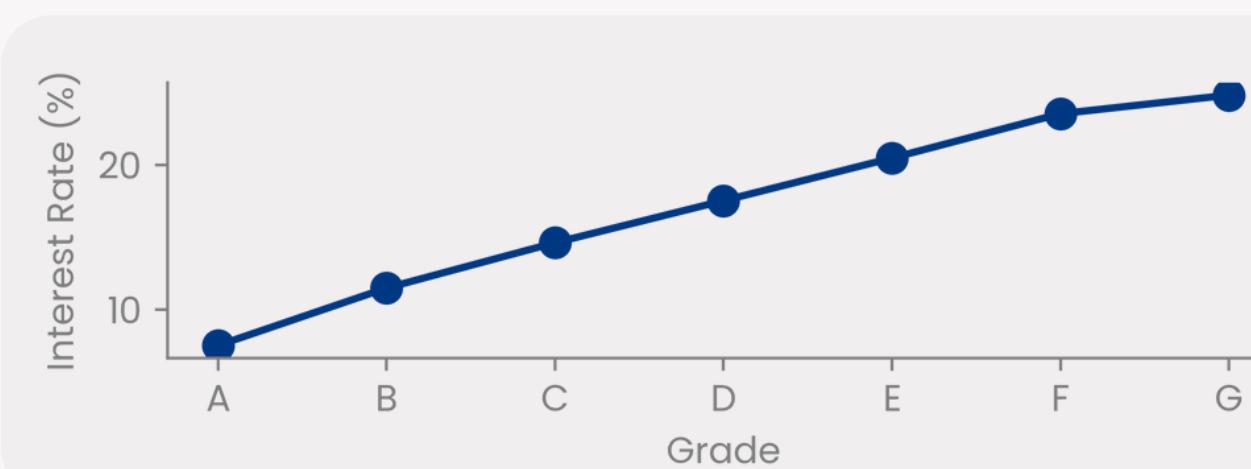
Restrict 60-month term to A-C grades or DTI <15%, or charge high premium for longer tenor risk.

# Portfolio Skewed Toward Mid-Grade Risk



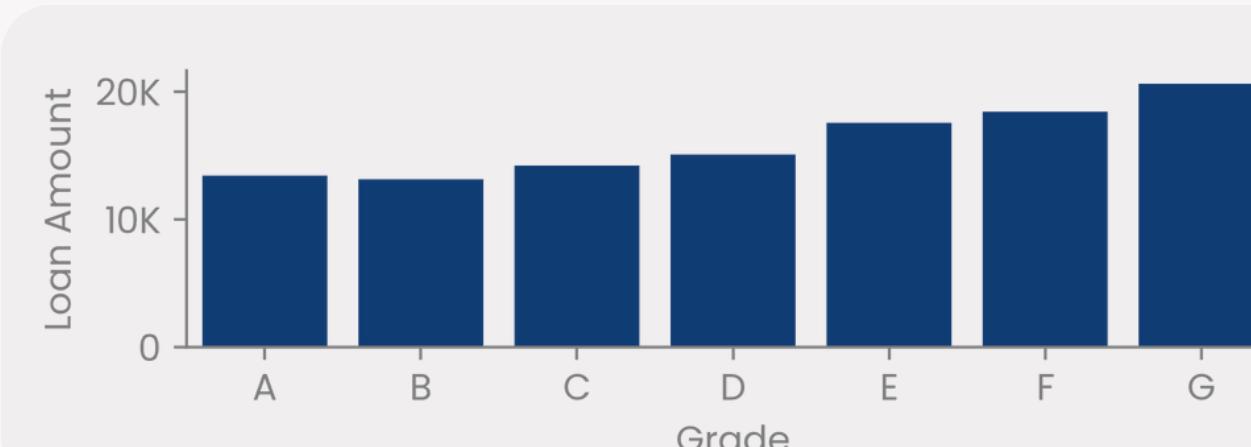
## Heavy Mid-Grade Concentration

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.



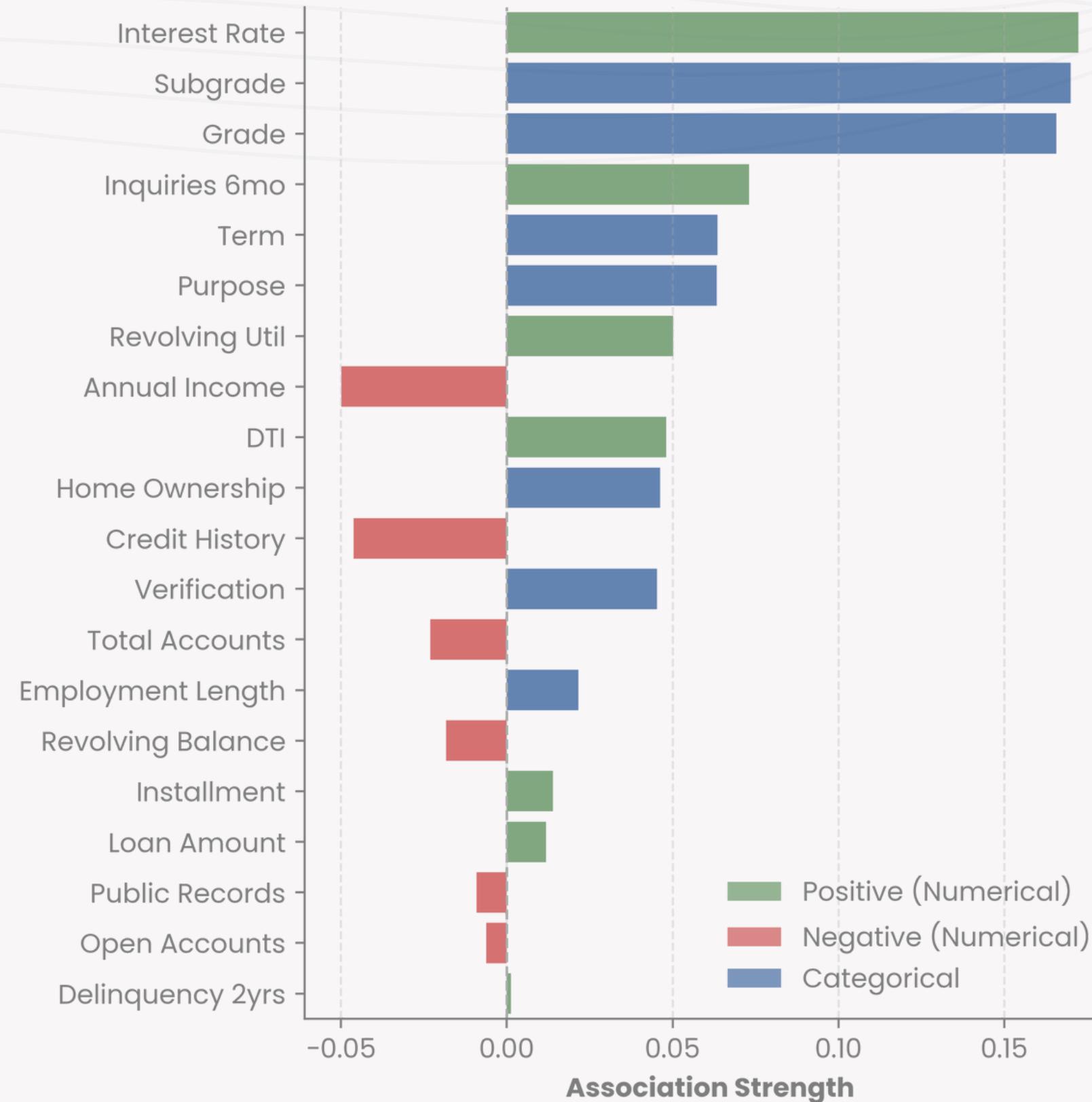
## Inverse Risk-Amount Relationship

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

# Feature Correlation Analysis



## Weak Linear Relationships Across All Features

Interest rate, sub\_grade, and grade 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, recent inquiries, loan purpose, DTI, and annual income. 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 (Chained Equations) + 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

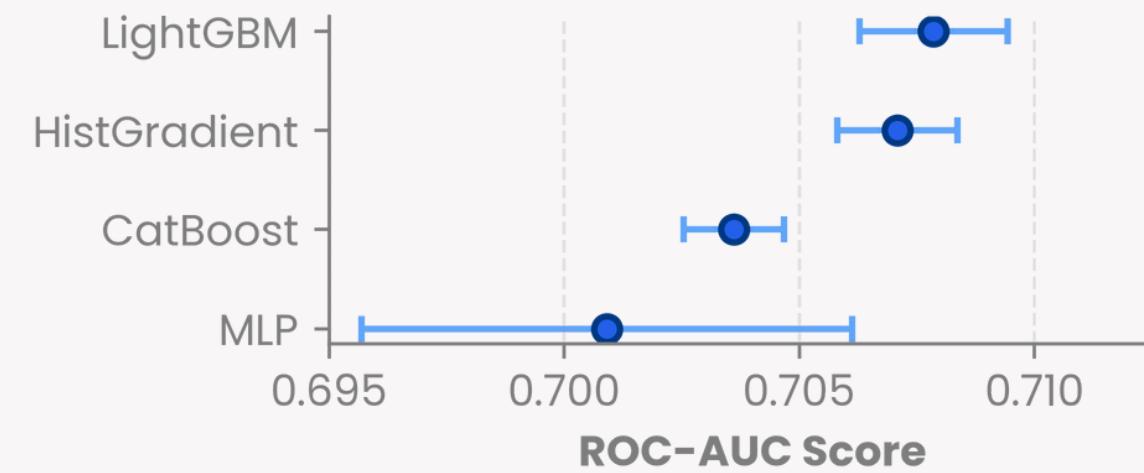
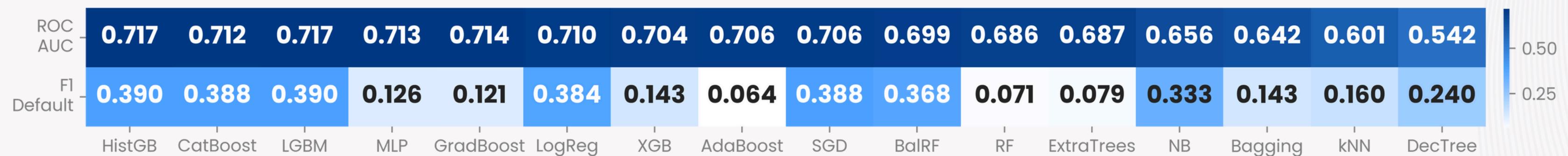
- L1 Regularization (Lasso) for linear models
- Feature Importance (Mean) for tree ensemble
- XGBoost Importance (Median) for boosting
- Variance Threshold (0.01) for distance-based



# Model Selection

## Initial Model Screening on Validation Set

Initial evaluation of **16 machine learning algorithms** showed ROC-AUC performance ranging from **0.542 to 0.717**, 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.

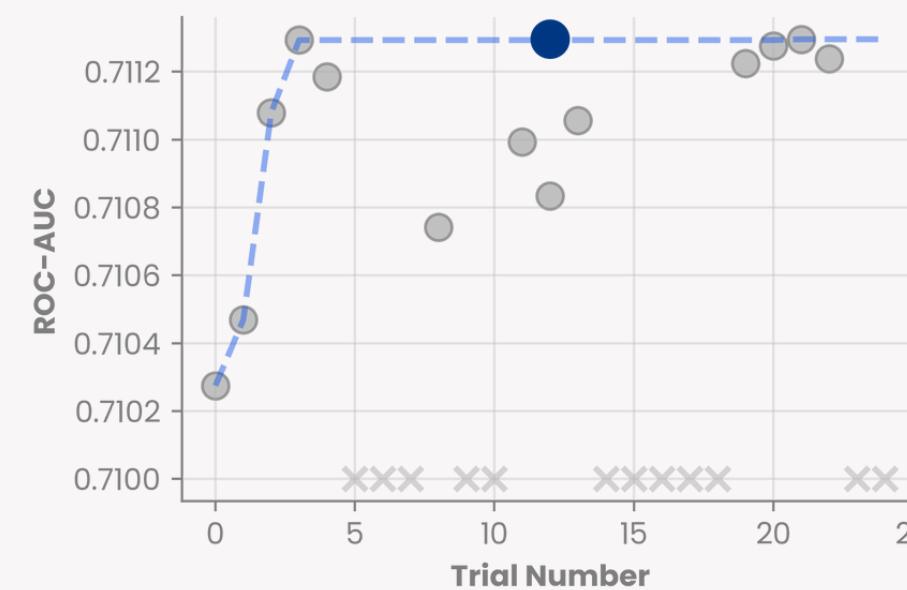


## Cross-Validation Results (5-Fold with 95% CI)

**LightGBM** was selected as the final model with a CV ROC-AUC of **0.7079** (highest) and a **stable confidence interval** (0.7063 - 0.7094). This model provides an **optimal balance** between **prediction performance** and **cross-fold consistency**, with **good recall** for positive case detection.

# Model Optimization

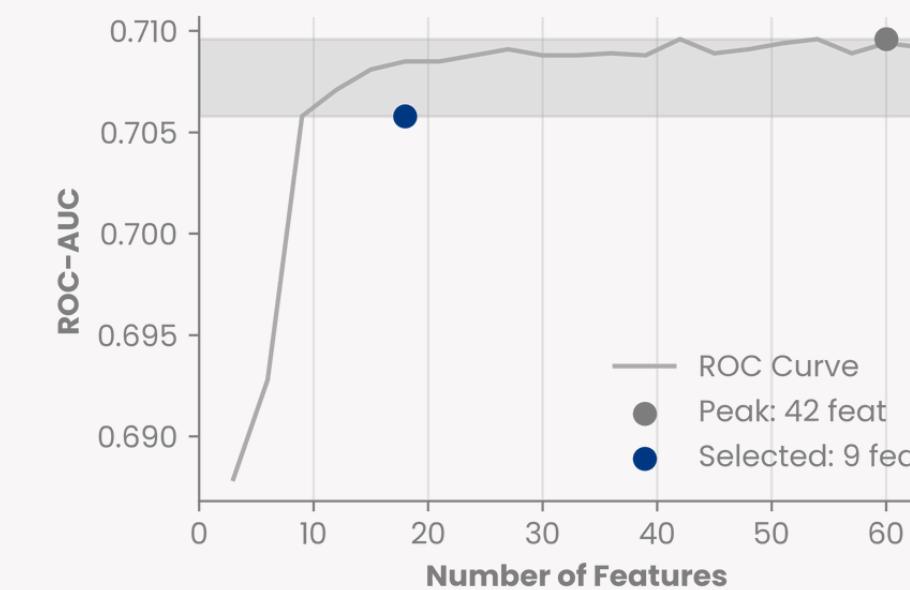
## Hyperparameter Tuning



**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**.

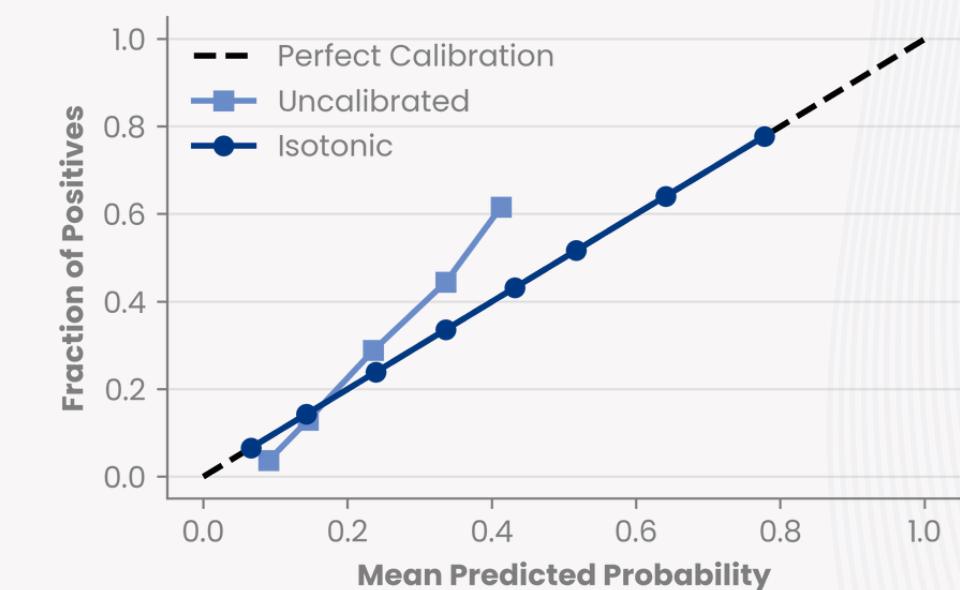
## Feature Optimization



**-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**  
**0.700**

Acceptable

**Brier Score**  
**0.128**

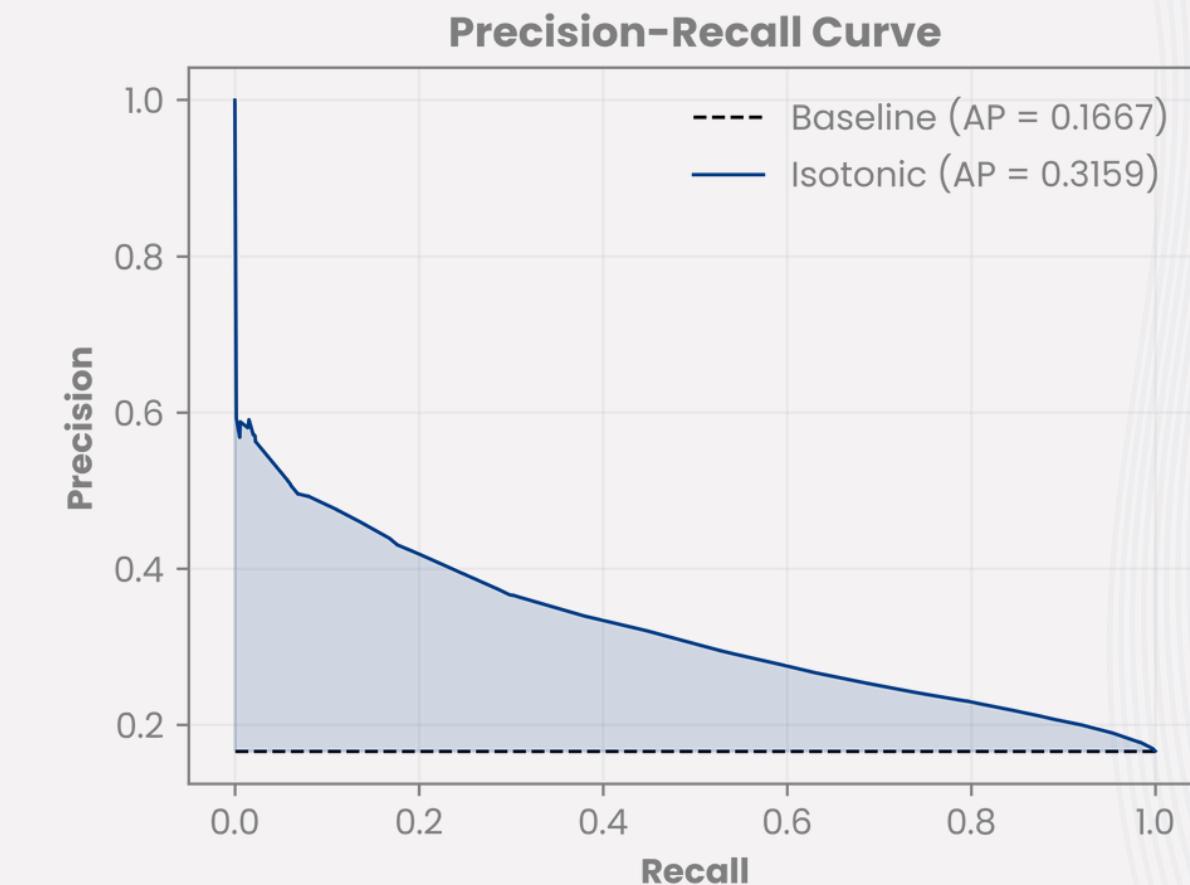
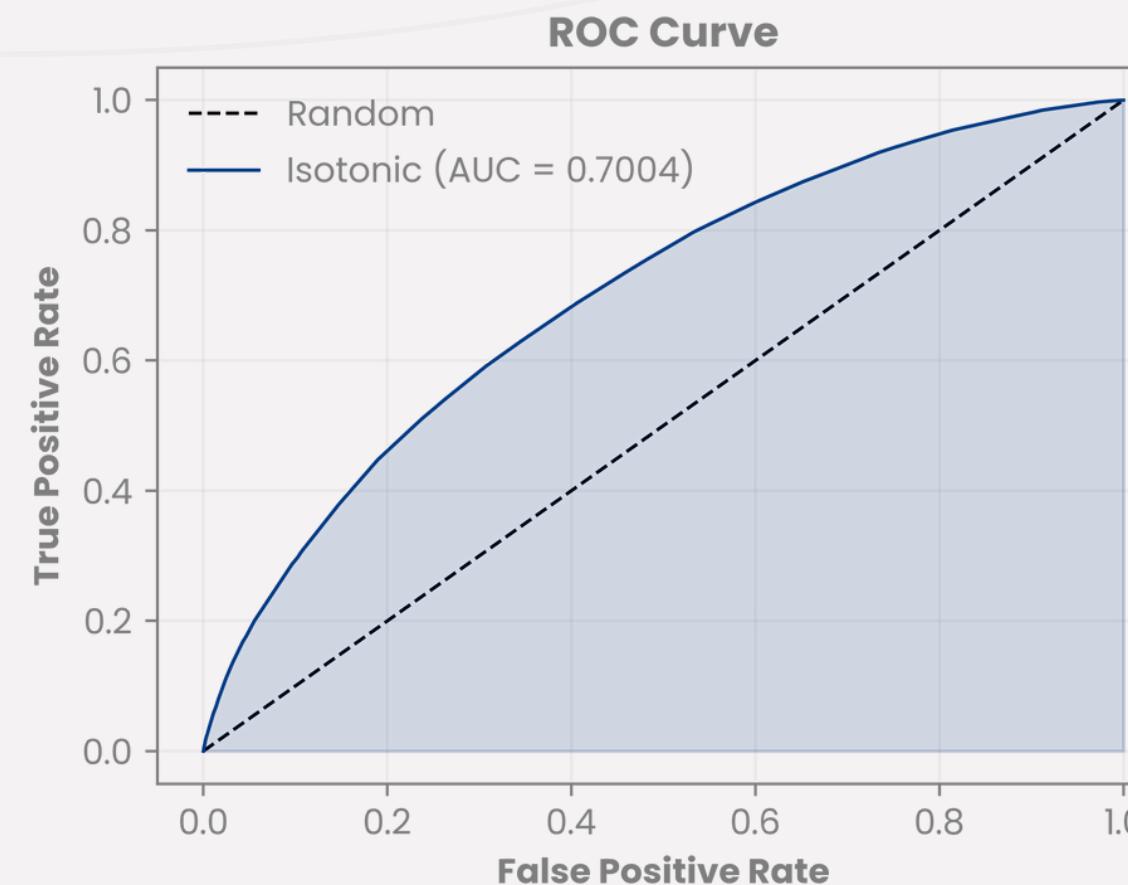
Excellent

**High Risk Precision**  
**42.8%**

Fair

**High Risk Recall**  
**20.8%**

Fair



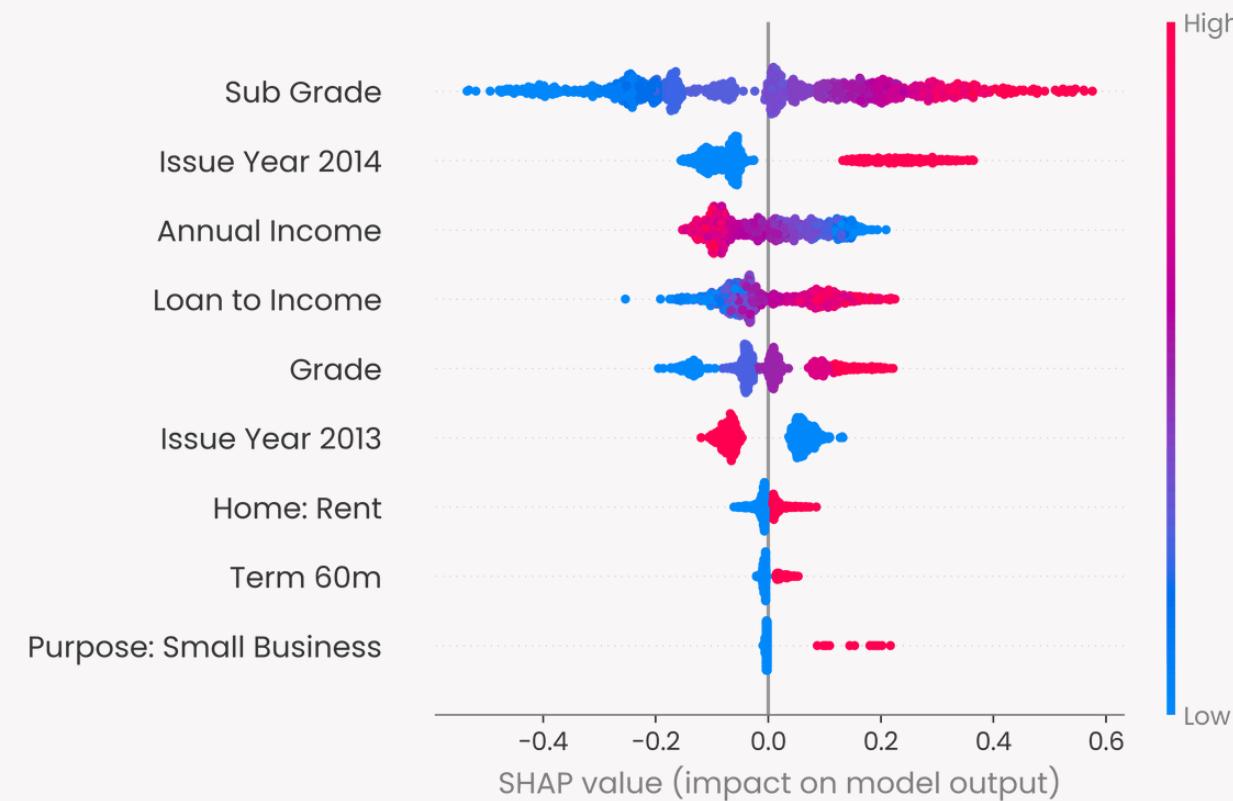
The model demonstrates **solid discriminative ability** with ROC-AUC of 0.708 and **well-calibrated probabilities** (Brier Score: 0.152). Overall accuracy reaches 78.98%. The lower PR-AUC (0.3977) reflects class imbalance, addressed through balanced class weights. **Tier segmentation** is applied for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

# Model Interpretability Overview

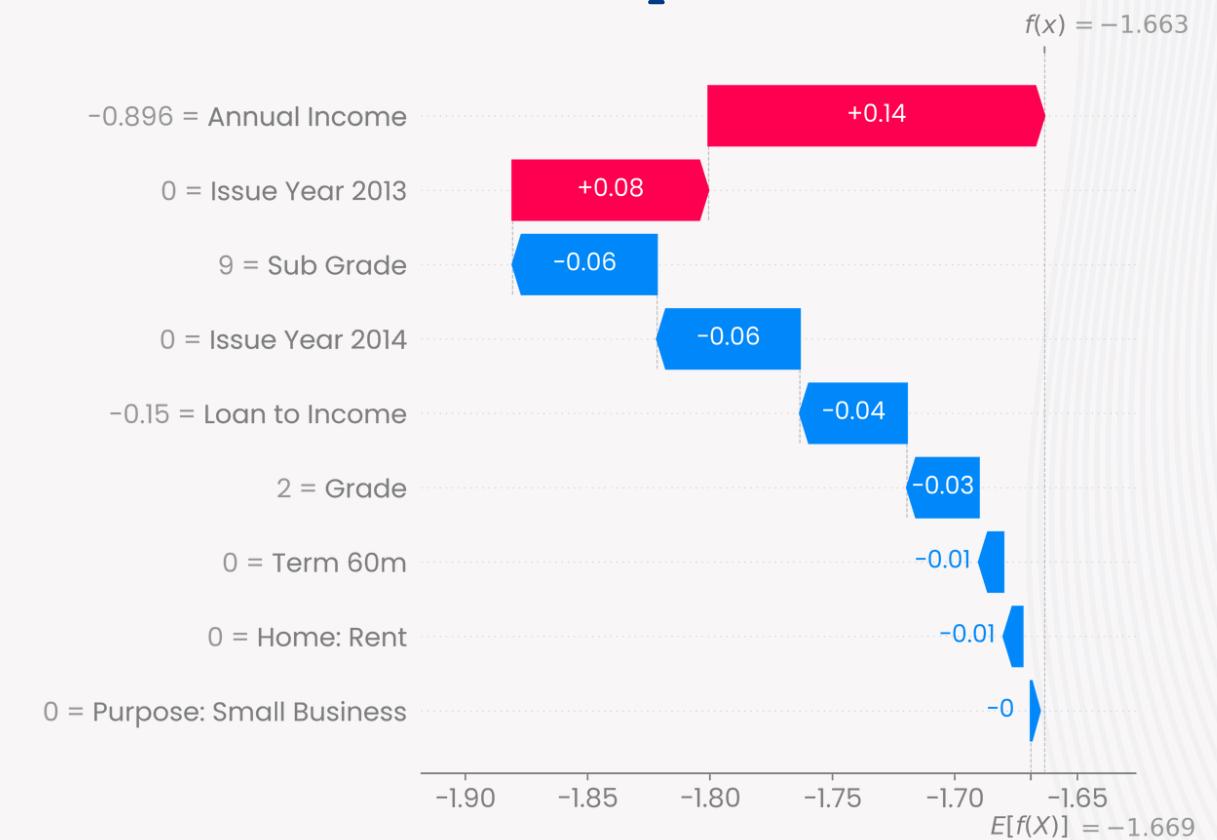
## Why SHAP Analysis?

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

## Global Feature Importance

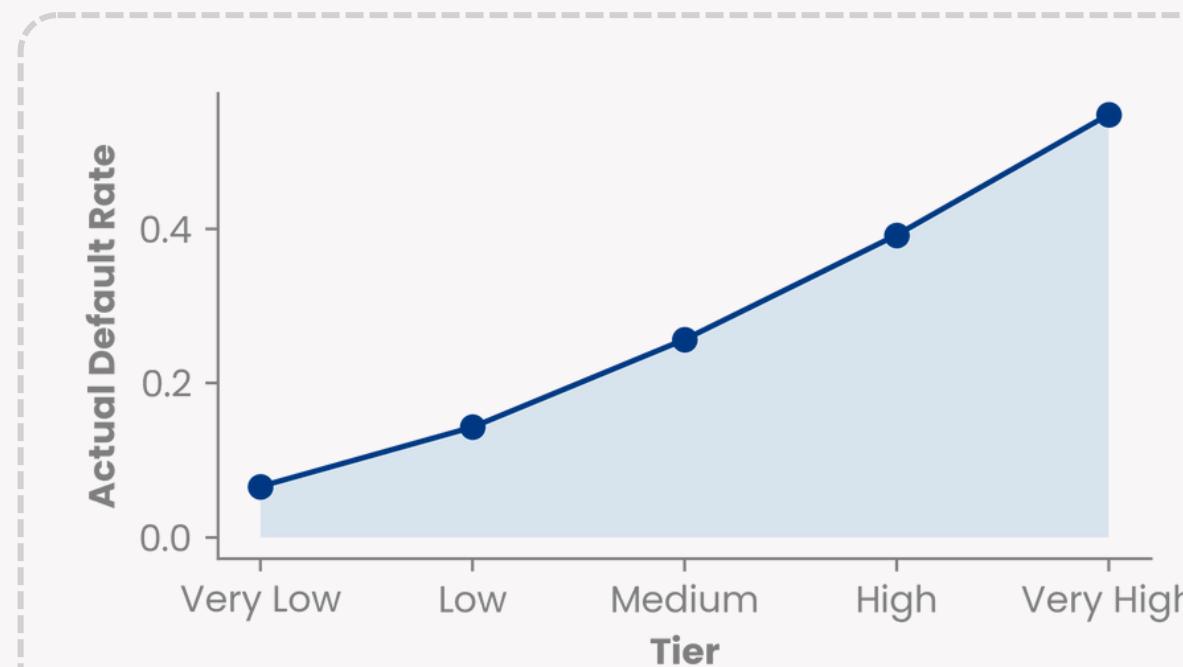
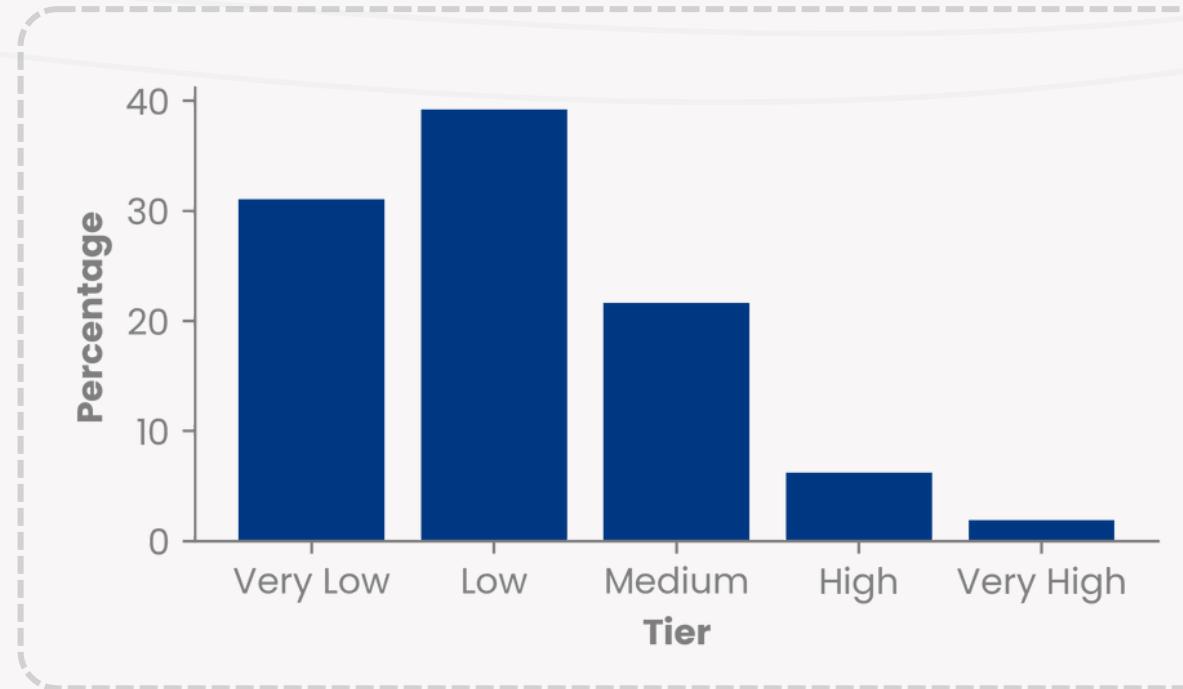


## Individual Case Example



SHAP analysis reveals how each feature contributes to predictions across the entire dataset and for individual cases. The **global view** shows **Sub Grade** and **Annual Income** as the **most influential features** with the **widest impact ranges**, while temporal factors like **Issue Year** show **moderate but consistent effects**. The **waterfall plot** demonstrates how **these features combine in practice**, Annual Income strongly increases this applicant's score, while Sub Grade and Issue Year counterbalance with negative contributions, resulting in the final prediction of -1.669.

# Risk Tier Segmentation



## Risk Tier Actions

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

	Threshold	Volume	Default Rate	Predicted Prob
<b>Very Low</b>	10%	15,187	6.6%	6.6%
<b>Low</b>	20%	19,182	14.3%	14.3%
<b>Medium</b>	32%	10,577	25.7%	25.7%
<b>High</b>	50%	3,038	39.2%	39.2%
<b>Very High</b>	100%	923	54.8%	54.8%

This method uses grid search to find optimal thresholds that maximize F-statistic (ANOVA) while ensuring monotonic default rate progression (6.6% → 54.8%) and minimum 0.5% volume per tier. After evaluating 1,000+ threshold combinations, the algorithm selects boundaries that provide the strongest statistical separation between risk tiers while maintaining business viability.

# Model Impact

\$10,1M

Losses Avoided

16.01%

Default Rate

26.07%

Model ROA

# Executive Summary

The lending company faces significant credit risk requiring an effective predictive model to mitigate financial exposure. A LightGBM-based credit risk model achieving 0.708 ROC-AUC was developed with a 5-tier risk segmentation system, enabling automated borrower classification and data-driven lending decisions.

## Critical Findings

- Mid-grade (B-C) dominates 56% while premium Grade A only 16%
- Q1 shows highest defaults, Q4 lowest across all regimes
- Grades D-E are sweet spot (14-18% default); F-G margins don't justify risk
- DTI <15% safest; DTI >35% outperforms 25-35% in stable periods
- 60-month terms default 41% higher than 36-month (12.3% vs 8.7%)

## Strategic Recommendations

- Increase Grade A-B from 16% to 25-30%
- Focus on D-E; exit or reprice F-G
- Auto-approve DTI <15%; add checks for DTI >35%
- Limit 60-month to Grade A-C or charge premium
- Target \$100K+ with stress tests for downturns
- Deploy 5-tier classification model for automated risk-based decisioning

# 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/)