UNSECURED LOAN RISK Modeling

aihack INDIA 2025

BEST TEAM RANKING : 4TH

From the shapes in nature, to signals in financial data— model: Pattern is the story, pattern is the solution."

Understanding The Core Challenge

Financial institutions must issue loans to grow –
but Unsecured Loans (no collateral) are high-risk decisions.

Banks can't perfectly identify who will repay vs default.
This causes two costly mistakes:

Approving a High-Risk Applicant	VS	Rejecting a Low-Risk Applicant

Loan Default	outcome	Missed Opportunity
Financial Loss	impact	Business Loss & Financial Exclusion

Impact of the Solution

For Borrowers

Borrowers get fair and faster credit access, even with limited history. The model reduces bias, ensuring decisions depend only on financial behavior. This boosts trust and transparency, improving overall customer experience.

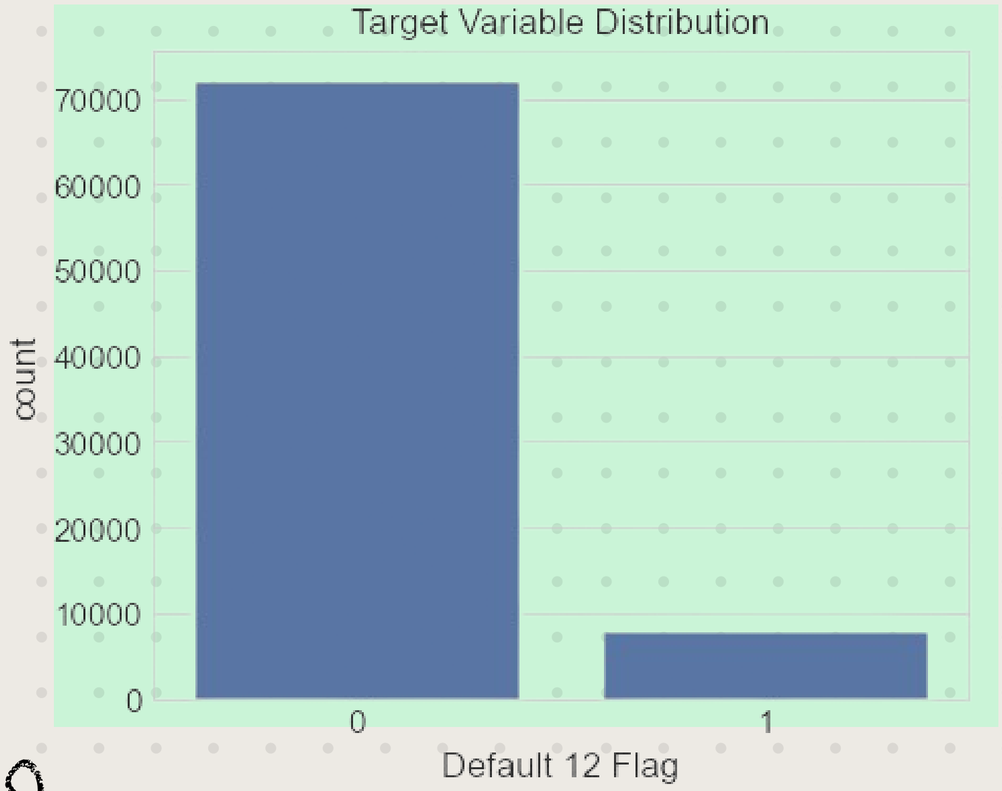
Financial Institutions (Lenders)

Our model helps banks cut loan defaults by 25–30% through early risk detection. It improves portfolio stability, saves ₹X lakhs–₹X crores in NPA losses, and reduces manual review time by 30–40%. Decisions become faster, smarter, and fully data-driven instead of gut-based.

For the Ecosystem

The system promotes financial inclusion and responsible lending, leading to fewer bad loans and stronger economic stability. It builds a smarter, fairer, and more resilient financial ecosystem.

Dataset Overview & Key Challenges



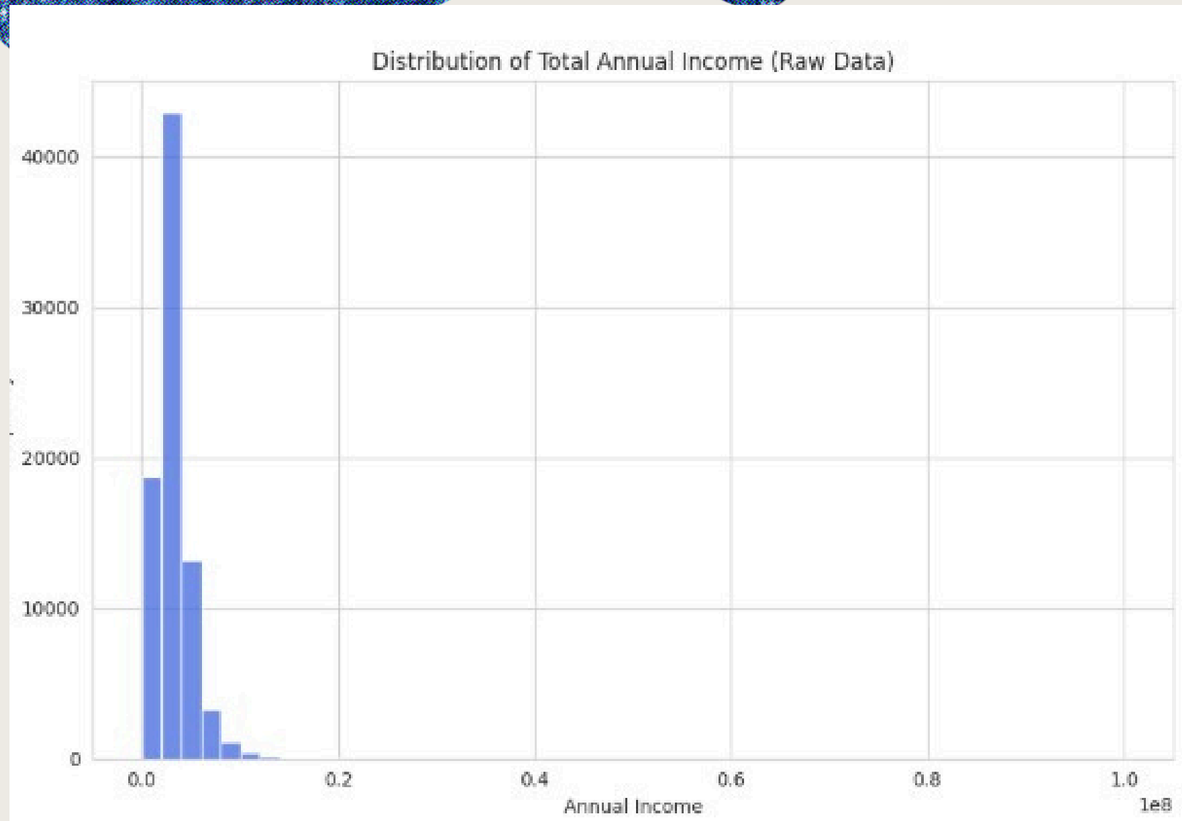
DATASET:
80K train, 20K test | Target: Default 12 Flag (1 = Defaulter) | Default rate ~9.9% → Imbalanced

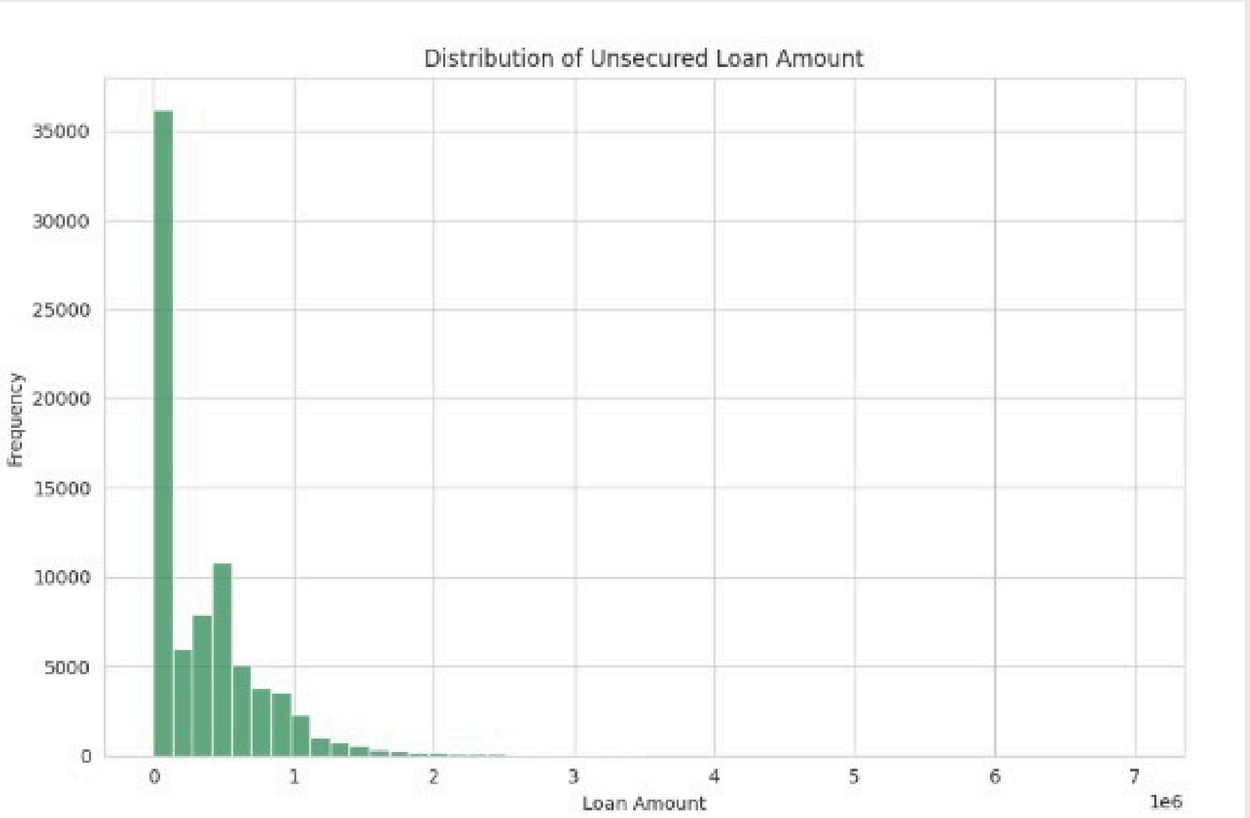
Feature Engineering:
Created 50+ new features (DTI ratios, honesty & stability scores) → AUC 0.59 → 0.675

Challenges:
Imbalance, hidden underreporting, overlapping borrower patterns, temporal drift.

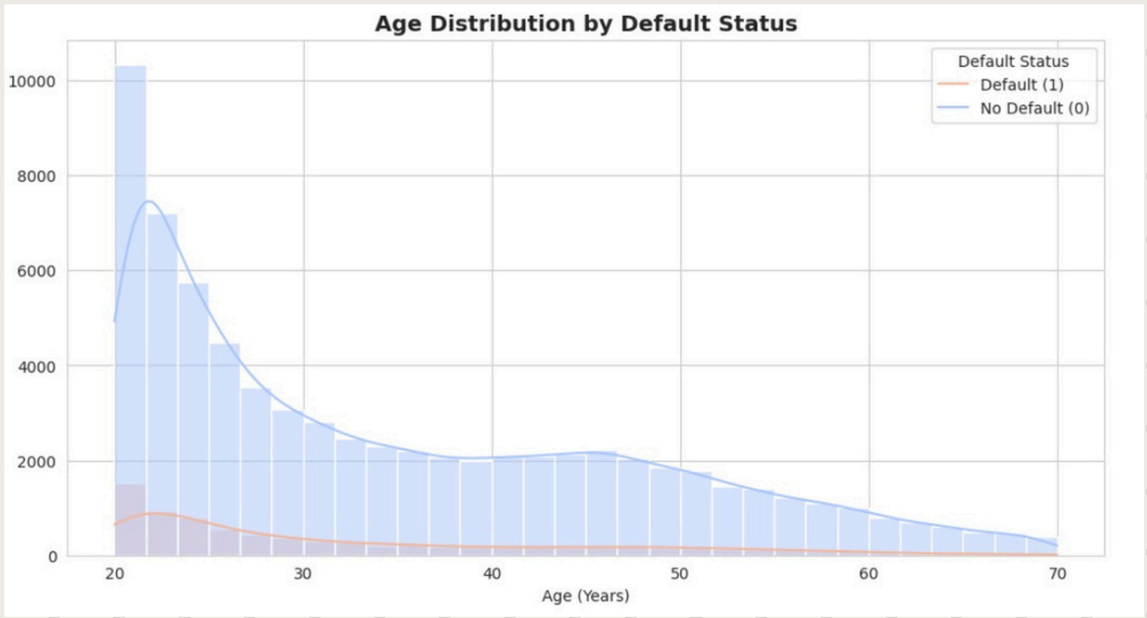
Insights:
Defaults linked to high loan intensity, multiple loans, renting, unstable jobs.

PRE PROCESSING





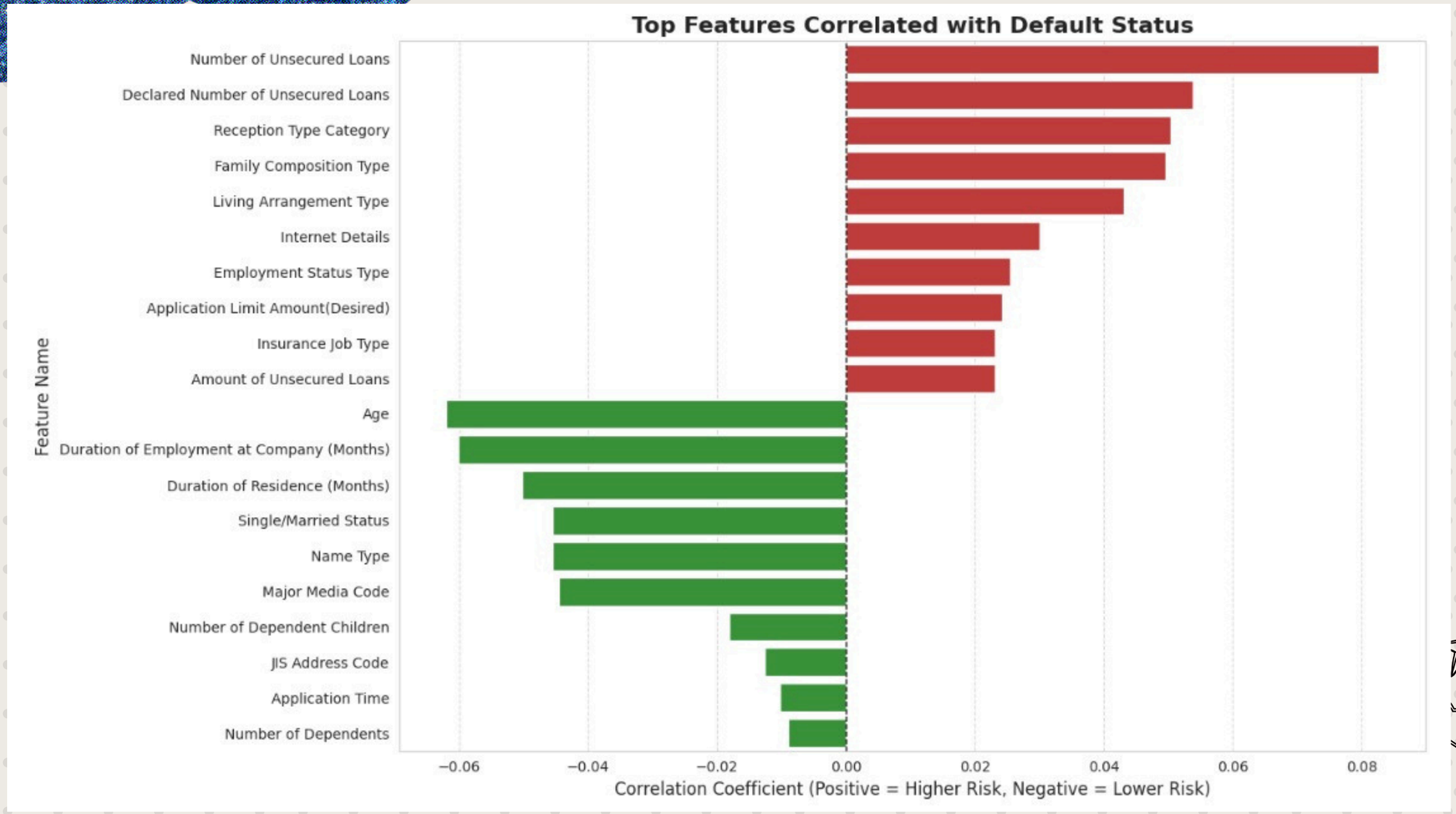
PRE PROCESSING



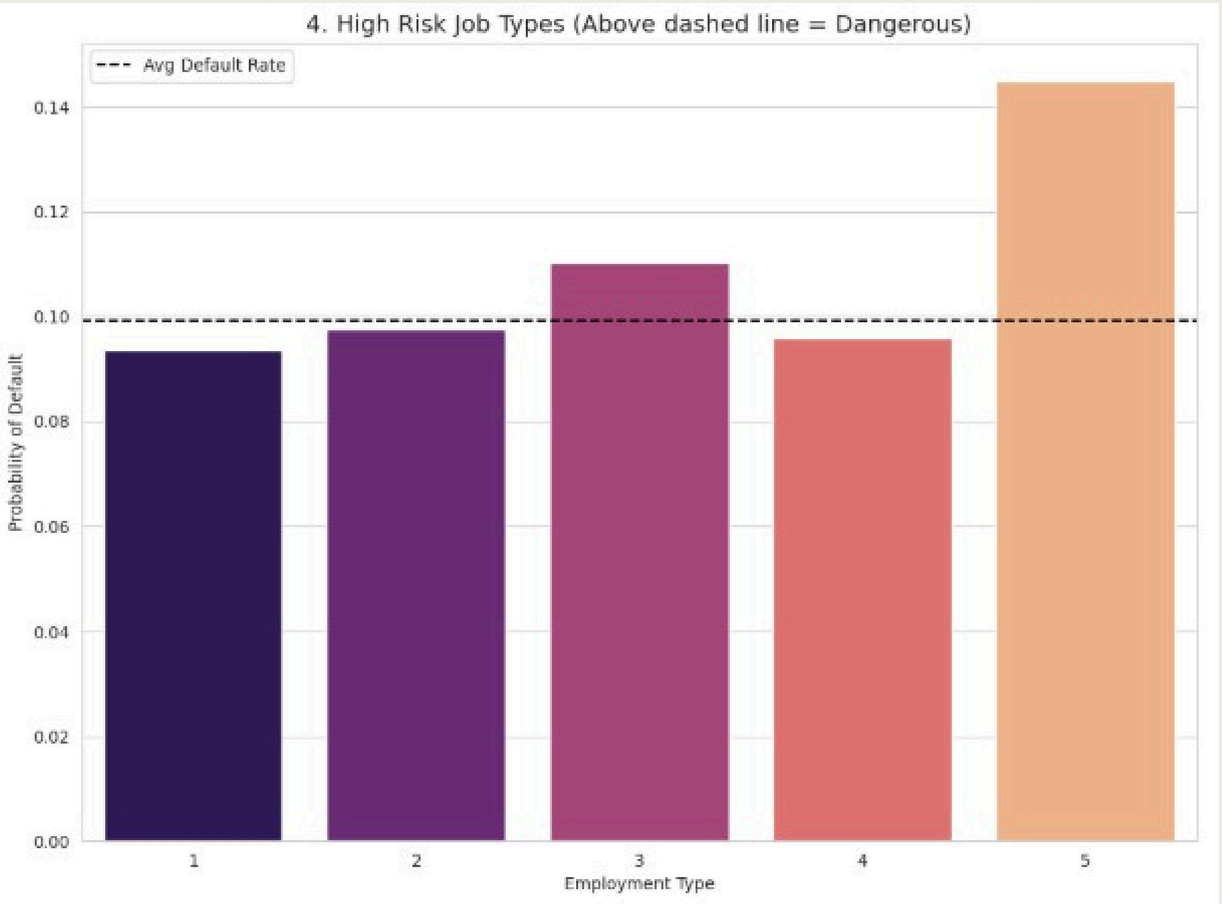

```
============================================================
📈 TOP 10 RISK FACTORS (Positive Correlation)
   (Higher values indicate higher likelihood of default)
============================================================
Number of Unsecured Loans                  : 0.0826
Declared Number of Unsecured Loans         : 0.0538
Reception Type Category                    : 0.0503
Family Composition Type                    : 0.0496
Living Arrangement Type                    : 0.0431
Internet Details                           : 0.0301
Employment Status Type                     : 0.0255
Application Limit Amount(Desired)          : 0.0242
Insurance Job Type                         : 0.0232
Amount of Unsecured Loans                  : 0.0231

============================================================
🛡️ TOP 10 PROTECTIVE FACTORS (Negative Correlation)
   (Higher values indicate lower likelihood of default)
============================================================
Age                                        : -0.0621
Duration of Employment at Company (Months) : -0.0602
Duration of Residence (Months)             : -0.0502
Single/Married Status                      : -0.0456
Name Type                                  : -0.0455
Major Media Code                           : -0.0445
Number of Dependent Children               : -0.0181
JIS Address Code                           : -0.0127
Application Time                           : -0.0103
Number of Dependents                       : -0.0090
```

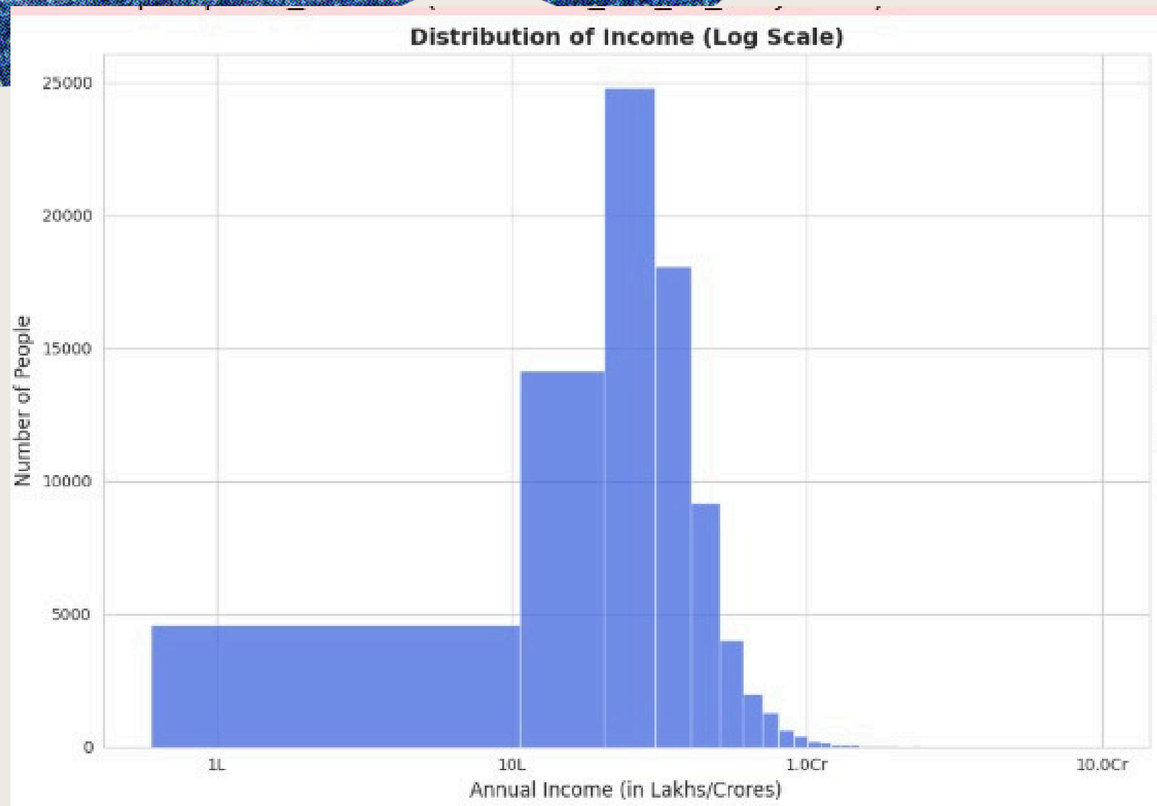
PRE PROCESSING

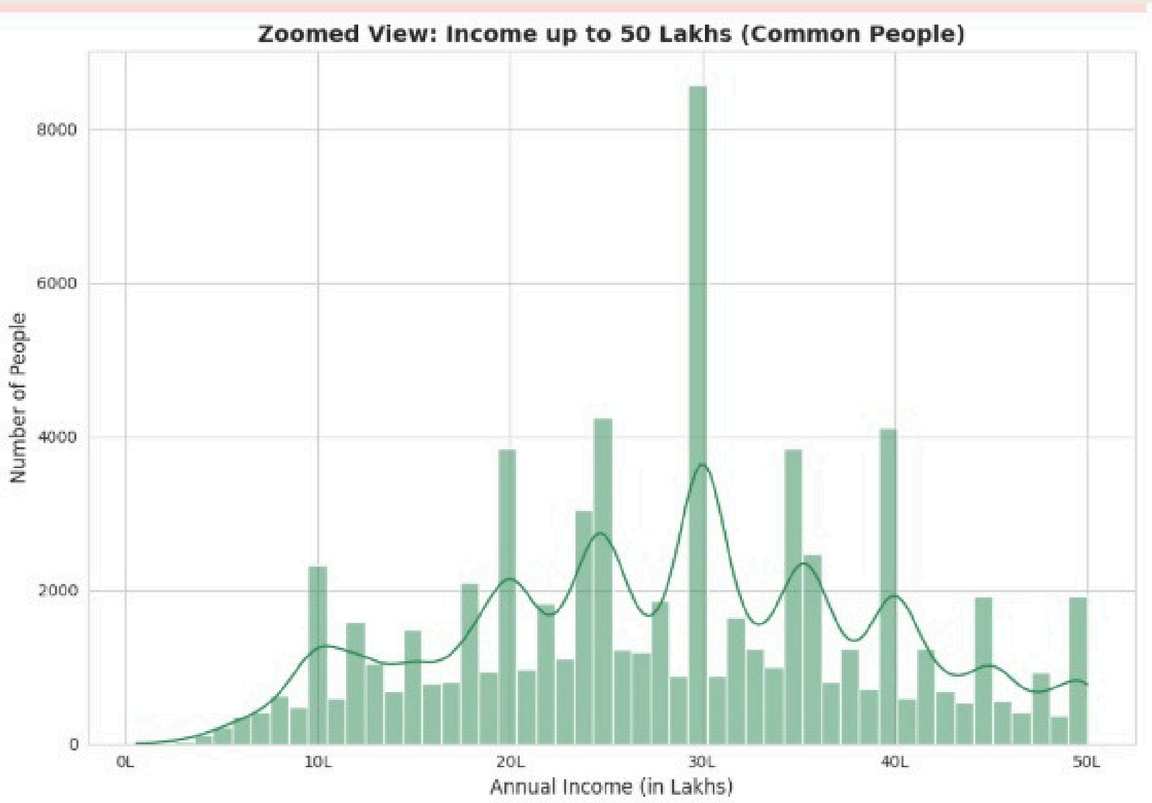


PRE PROCESSING

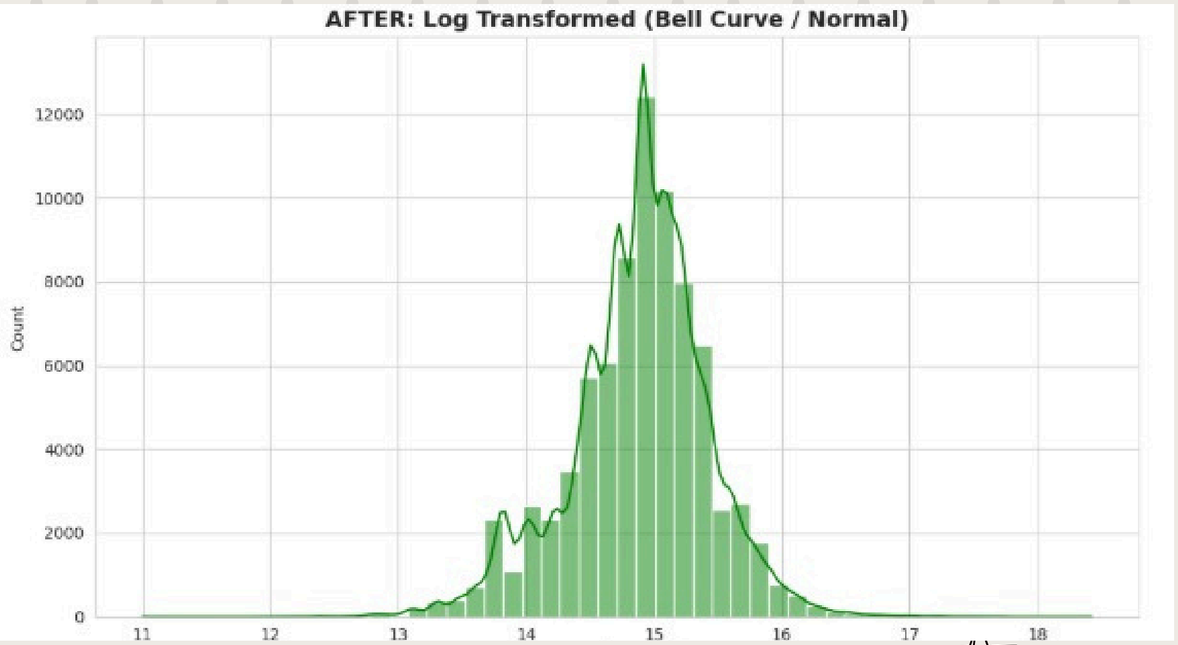


PRE PROCESSING





PRE PROCESSING



Risk Factor Analysis

- During the preprocessing stage, I conducted a detailed risk factor analysis to understand which borrower characteristics were linked with higher or lower default probability.
- employment types showed significantly higher default rates (often above 15%), while older applicants demonstrated much lower risk.
- Stability factors such as longer residence duration and longer employment duration also appeared as protective indicators. Additionally, applicants with multiple unsecured loans consistently showed higher default likelihood.

Outlier Handling & Numerical Stability

- Financial variables in the dataset—such as annual income, loan amount, rent burden, and number of loans—showed heavy skewness and extreme outliers.
- To prevent the models from being biased by these extreme values, I applied log transformations and conceptual winsorization logic to reduce their impact.
- These stability adjustments helped smooth the distributions and made the dataset more suitable for robust model training.

Phase 4 – Execution & Production

DATA SUMMARY & SETUP

- Dataset: 80K train, 20K test, target imbalance at 9.9%.
- Defined ~120 engineered features for stability & interpretability.

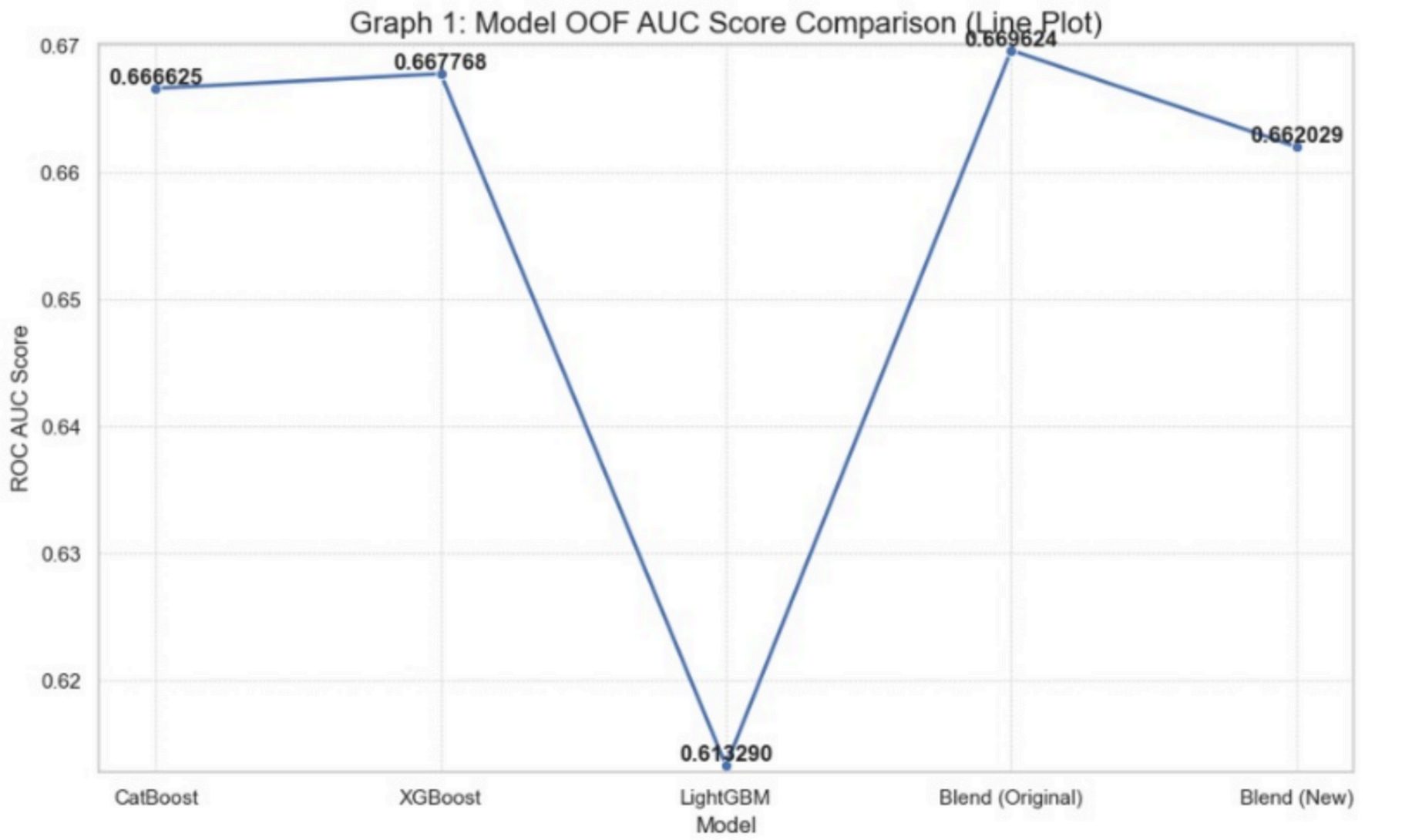
MODEL TRAINING PROCESS

- Used 5-Fold Stratified Cross-Validation for reliable results.
- Trained CatBoost and XGBoost with strong regularization to avoid overfitting.

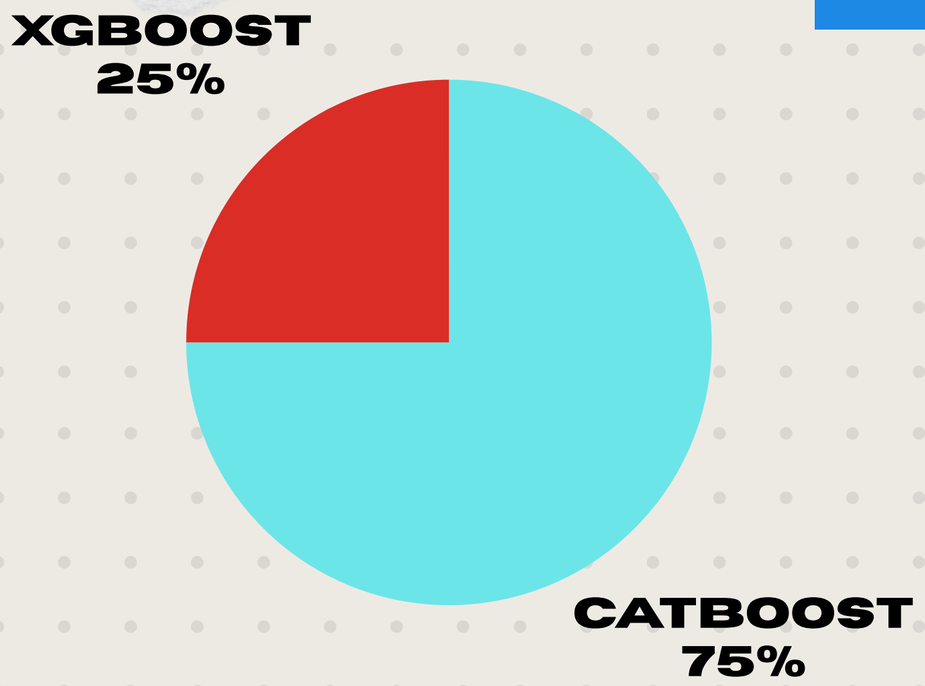
MODEL PERFORMANCE (PER FOLD)

- CatBoost AUC range: 0.666 – 0.680 across folds.
- XGBoost AUC range: 0.666 – 0.678, consistent and stable

Different Model Comparison

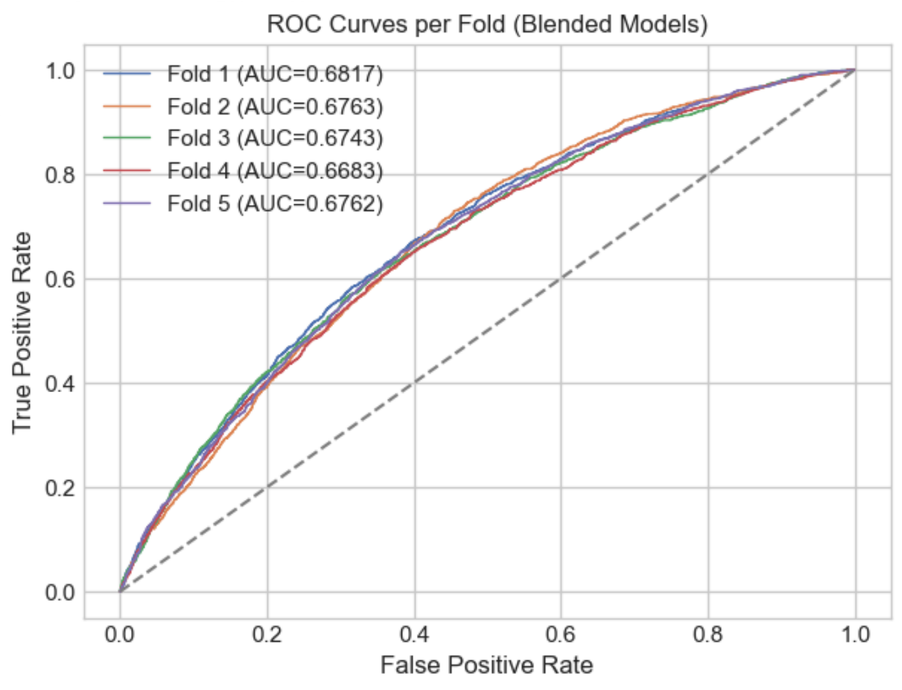


Final Score



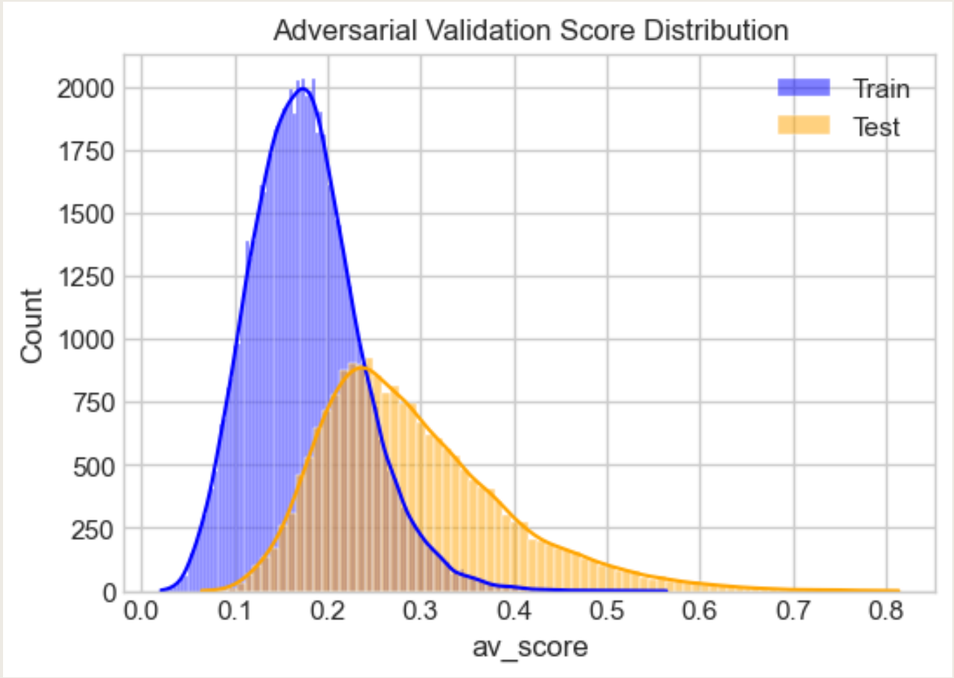
- Combined predictions: 75% CatBoost + 25% XGBoost.
- Achieved Final AUC = 0.677890, marking a +0.08 improvement over baseline.

trustworthiness + reliability



ROC-AUC

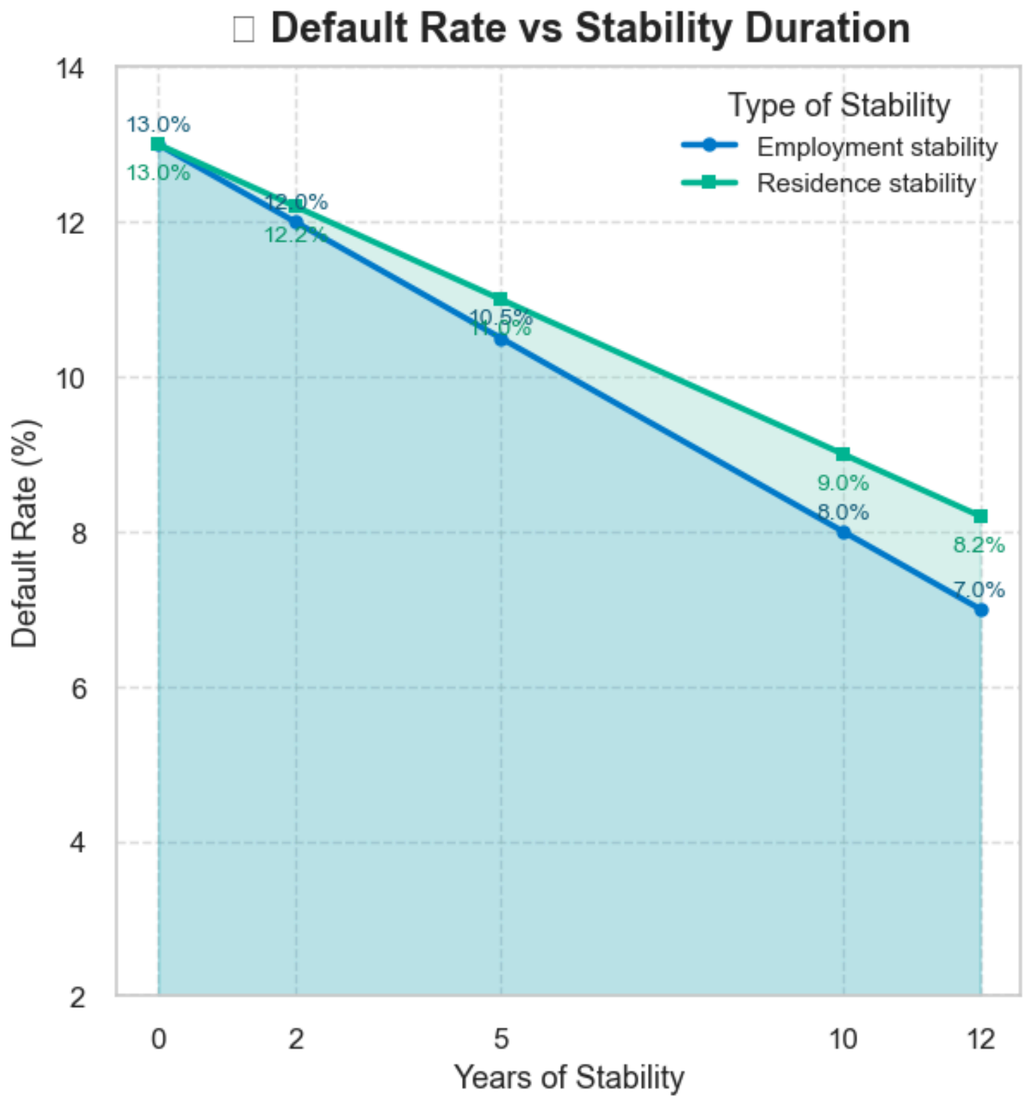
- AUC = 0.675 ± 0.004 → Stable and consistent performance across folds.
- Shows the model generalizes well, with no overfitting and balanced predictive power.



VALIDATION SCORE

- AV AUC ≈ 0.60 → Train and test distributions are well aligned.
- Minimal data drift, no leakage – ensures reliable generalization on unseen data.

Insight Discovery



Default Rate Estimation Formula

$$\text{Default Rate } (\%) = 13 - 0.5 \times E - 0.4 \times R$$

E: Employment Years

R: Residence Years

13% = Baseline default rate for new applicants

Each additional year of employment reduces default risk by ~1 percentage point

Two additional years of residence reduce default risk by ~1 percentage point
-0.4 × Residence years lowers it further

Stability in job & home can cut risk by 70–75%

Example Predictions:

Employ. (yrs)	Residence (yrs)	Pred. Default (%)
0	0	13.0
2	1	11.6
5	3	9.3
10	5	6.0
12	10	3.0

Combined job & residence stability sharply cuts risk.
Easy, interpretable predictors for repayments.

Insight Discovery

HONEST BUT HIGH DTI

VS

DISHONEST BUT LOW DTI

- Declares all loans clearly – default only ~12%.
- Looks safe, but defaults ~15% due to hidden debt.
- Honesty offsets high DTI – 2-3% lower default vs dishonest peers.
- Low DTI hides unpaid loans – risk stays high.
- Clear repayment pattern, easier to trust.
- Models misjudge risk – defaults come unexpectedly.
- Honest borrowers stay reliable long-term.
- Low DTI but dishonest = poor future credibility.

Business view

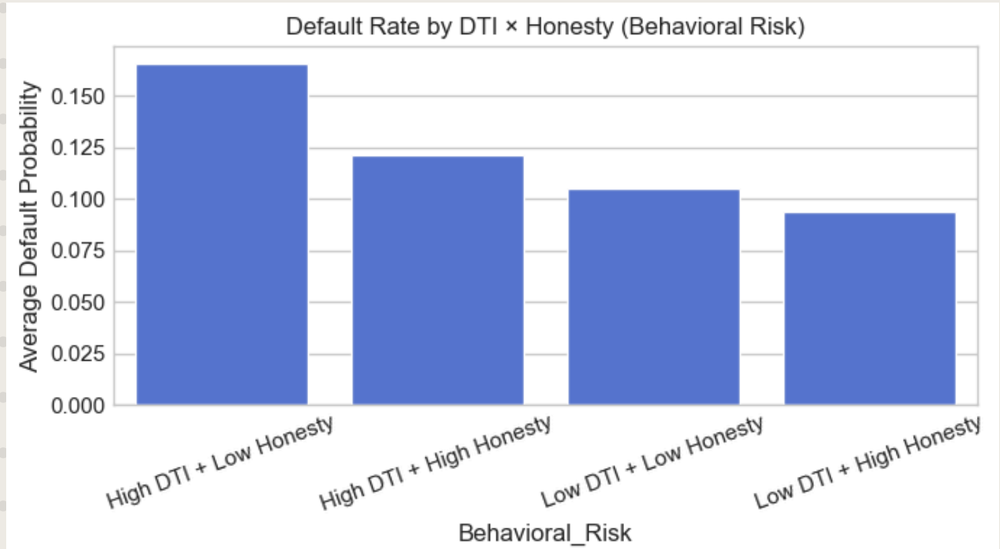
TOP 5 FACTORS DEFINING SAFE VS RISKY BORROWERS

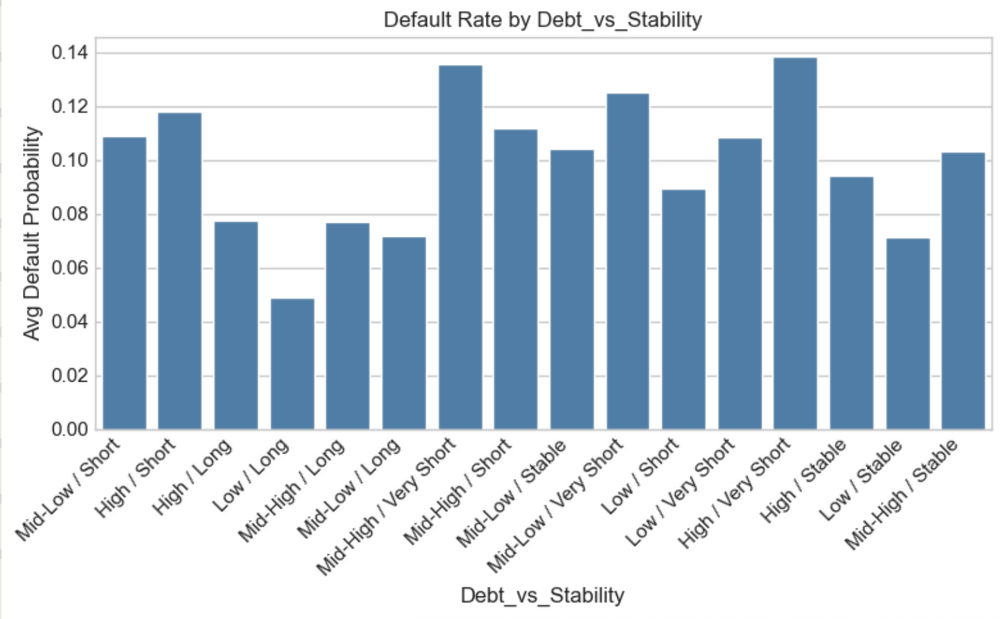
SAFE

- Long job + Homeowner
- Govt / Financial / Education sector
- Married, 2-3 dependents
- Low-mid DTI & steady income

RISKY

- Short job + No home
- Printing / Construction / Real Estate
- Single / 5+ dependents
- High DTI / loan overload






thank
you