Home Credit Default Risk: Final

Predicting Loan Default with Machine Learning

Group 1 | DATA 230 | Dec 03, 2025
Members: Liana Pakingan, Louisa Stumpf, Xu Wang, Yuyao Ding

Content

01 Brief Recap

Dataset Overview

Source

Kaggle Home Credit Default Risk

Size

~ 300k rows, 120+ features

Structure

Main application + 6 linked tables

Target

TARGET (1= default, 0 = non-default)

Features

Numeric, categorical, time, financial

Complexity

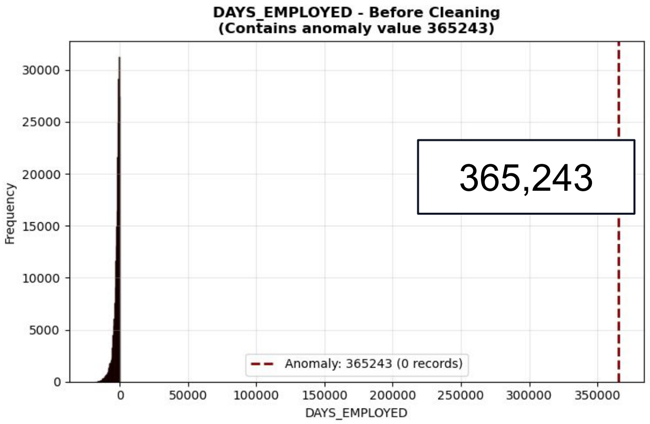
Missing values, outliers, imbalance(8%)

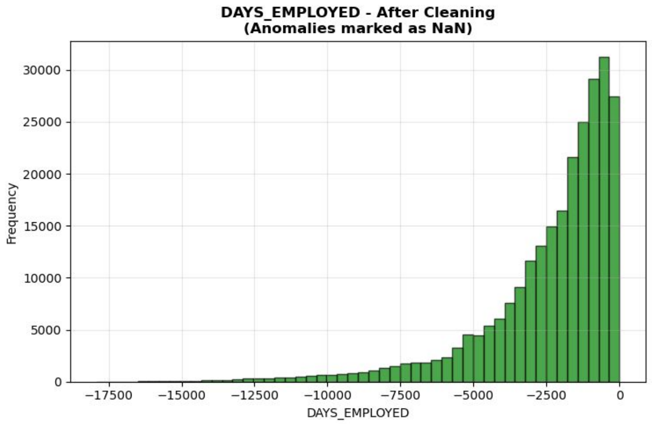
Key Data Processing–Clean, consistent data foundation for modeling

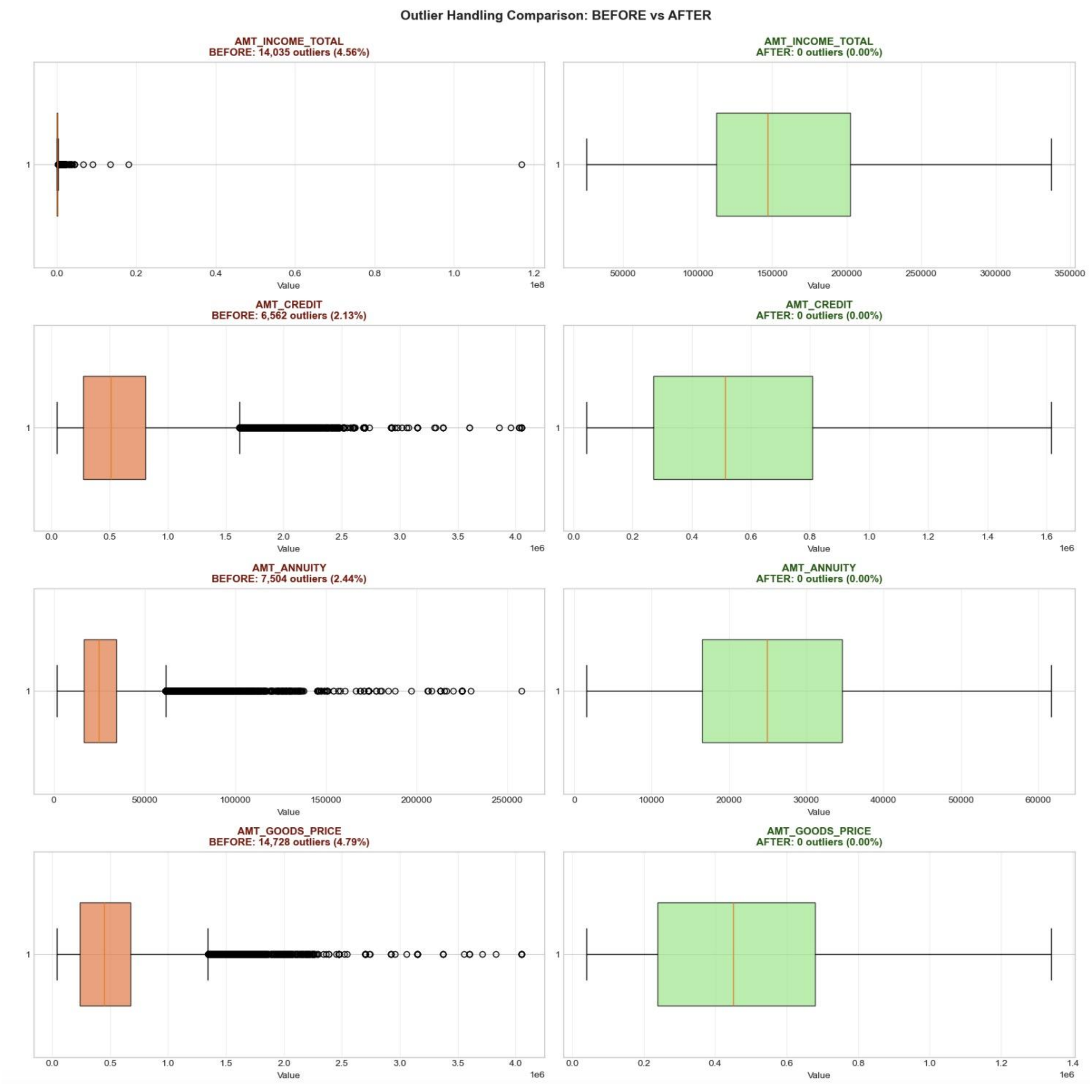
Missing Values
- 70% threshold, median/mode imputation → 22.65% to 0% missing

Outliers
- IQR-based capping with business constraints → reduced 4.5% outliers to 0%

Data Integration
- Merged 7 tables → 122 to 185 features







EDA Key Findings

Class Imblance Discovery

- Only 8.07% defaults → informed metric choice (ROC-AUC over accuracy)

Strongest Predictors(Ext_source) Identified

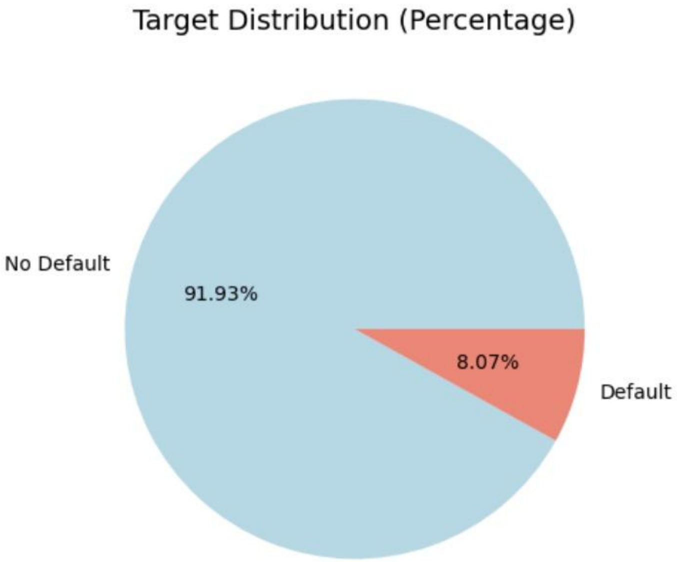
- Correlation with TARGET: -0.16 to -0.18

Feature Relationship

- Age/employment show moderate predictive power

Modeling Implications

- Use balanced class weights
- Prioritize EXT_SOURCE features
- Engineer ratio features



15 new features created across 5 categories

Category	Count	Examples	Purpose
Financial Ratios	4	CREDIT_INCOME_RATIO	Relative burden
Temporal	3	AGE_YEARS, EMPLOYED_YEARS	Stability
EXT_SOURCE Agg	2	EXT_SOURCE_MEAN	Combined signals
Per Capita Income	2	INCOME_PER_FAMILY	Household burden
Log Transform	4	AMT_INCOME_TOTAL_LOG	Handle skewness

02 ML Results

Model Performance Overview

Model Compared

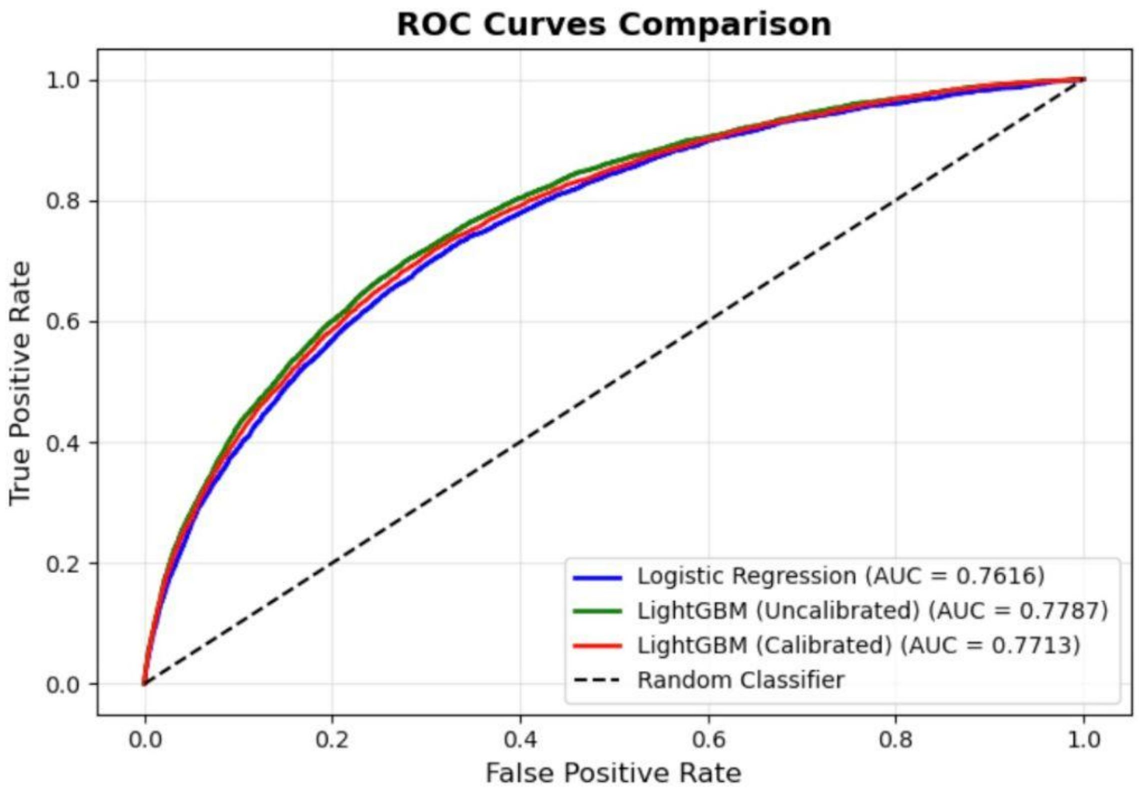
- ❖ **Logistic Regression (Baseline):** Interpretable, well-calibrated probabilities
- ❖ **LightGBM (Primary):** State-of-the-art gradient boosting for tabular data

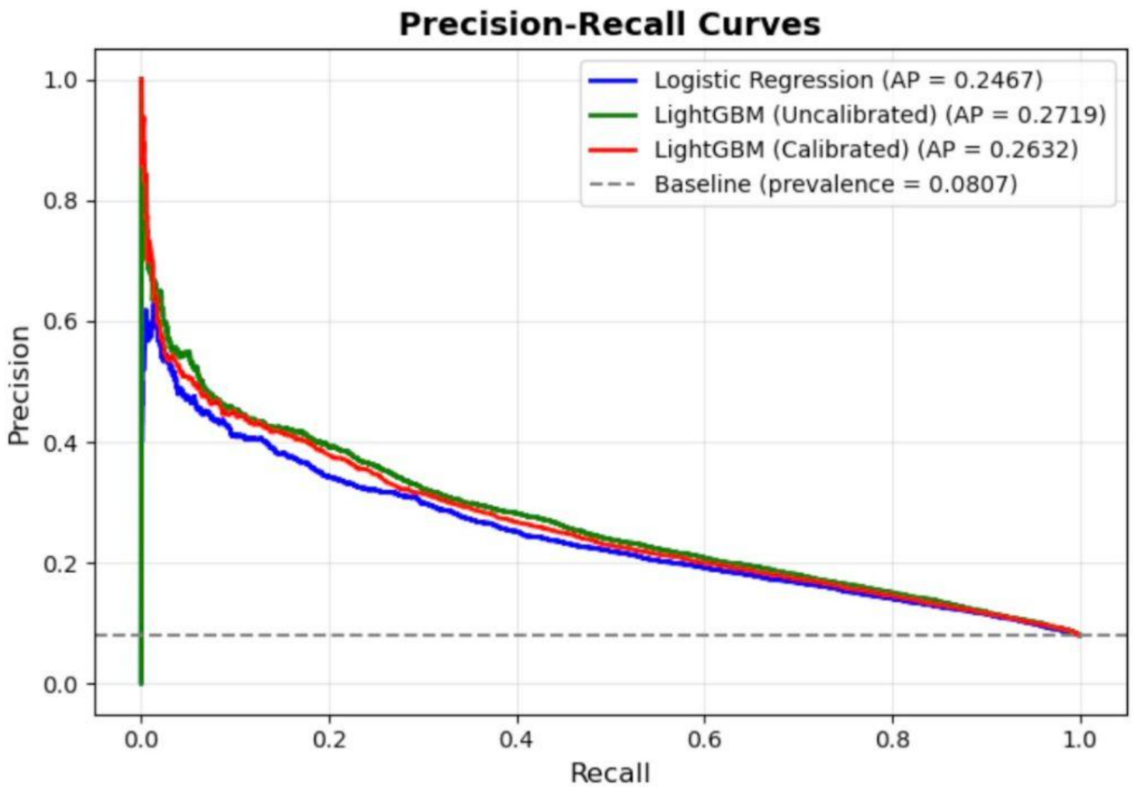
Model	ROC-AUC	Avg Precision	Brier Score	Log Loss
Logistic Regression	0.7616	0.2467	0.1983	0.5815
LightGBM	0.7787	0.2719	0.1771	0.5270
LightGBM (Calibrated)	0.7713	0.2632	0.0669⭐	0.2419⭐

Key Takeways:

- **LightGBM outperforms baseline across all metrics**
- **Calibration dramatically improves probability reliability (Brier: 0.1771→0.0669)**
- **Slight AUC trade-off acceptable for business-ready probabilities**

ROC & Precision-Recall Curves

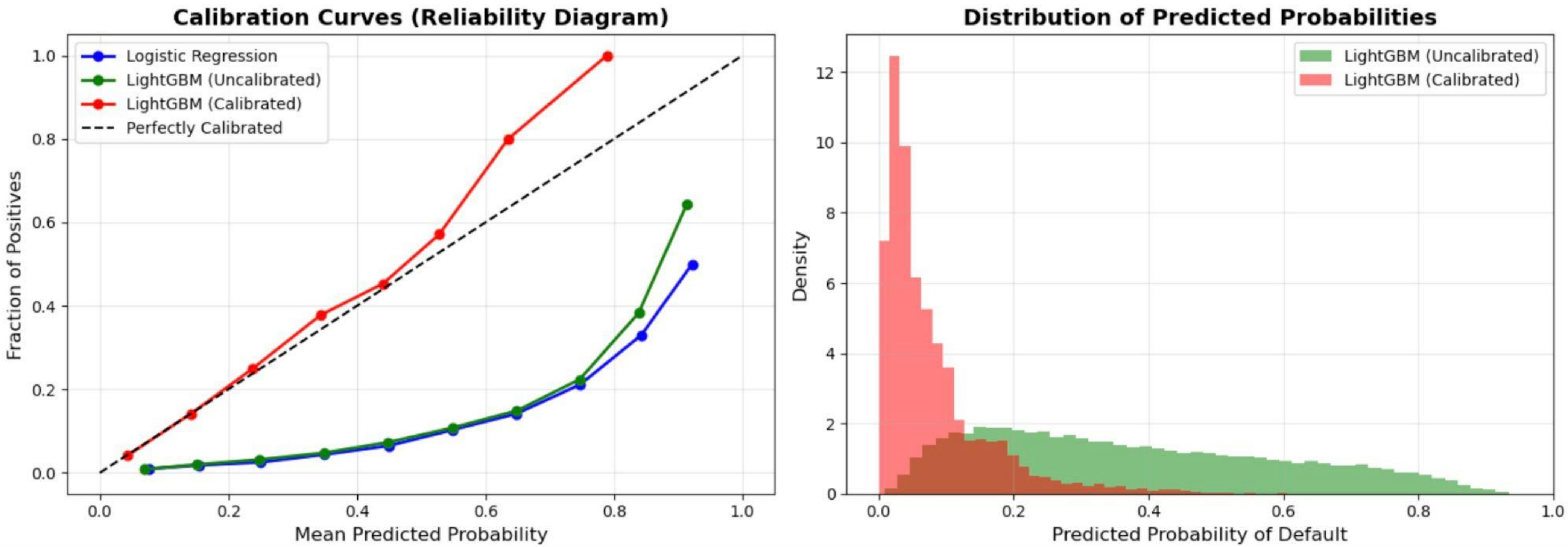




Interpretation:

- ROC Analysis: LightGBM achieves superior discrimination (AUC 0.7787 vs 0.7616)
- PR Analysis: More informative for imbalanced data (8% default rate)
  - LightGBM AP=0.2719 → better at identifying defaults with fewer false positives
  - Baseline AP=0.0807 (prevalence) → model adds significant value
- Calibration impact: Slight AUC decrease (-0.74%) acceptable for reliable probabilities

Probability Calibration



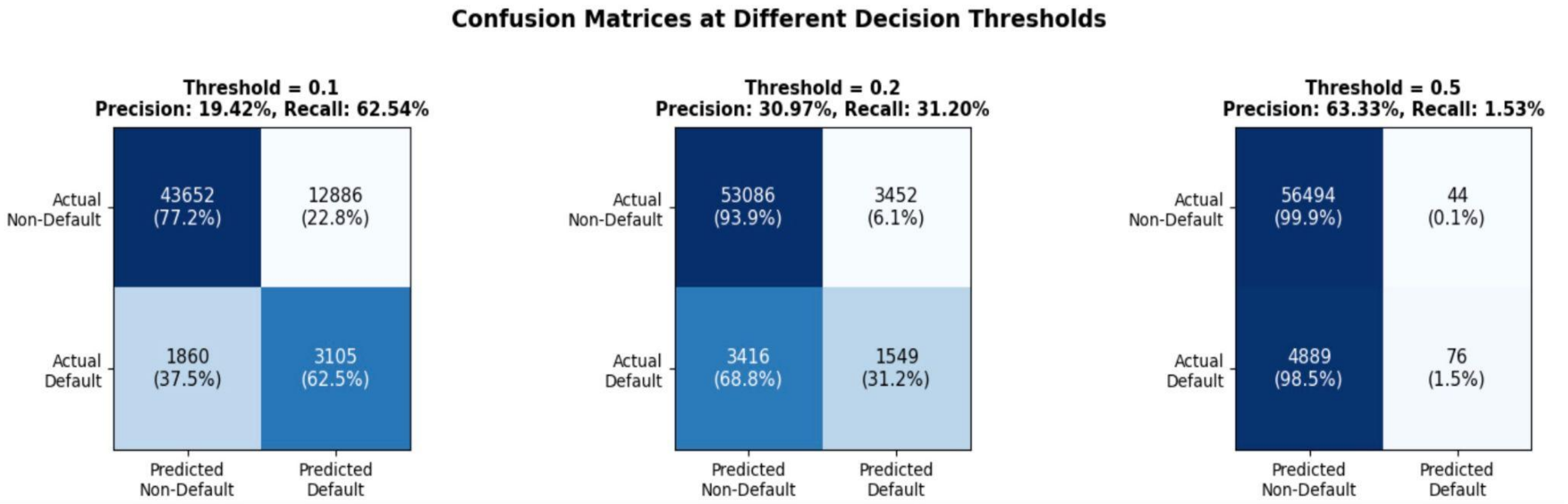
Why Calibration Matters:

- ❖ Raw LightGBM predictions underestimate default probabilities
- ❖ Uncalibrated: predicts 20% → actual 35% default
- ❖ Calibrated: predicts 20% → actual ~20% default ✅

Business Impact:

- ❖ Risk-based pricing: Accurate PD needed for interest rate setting
- ❖ Portfolio management: Expected Loss = PD × LGD × EAD

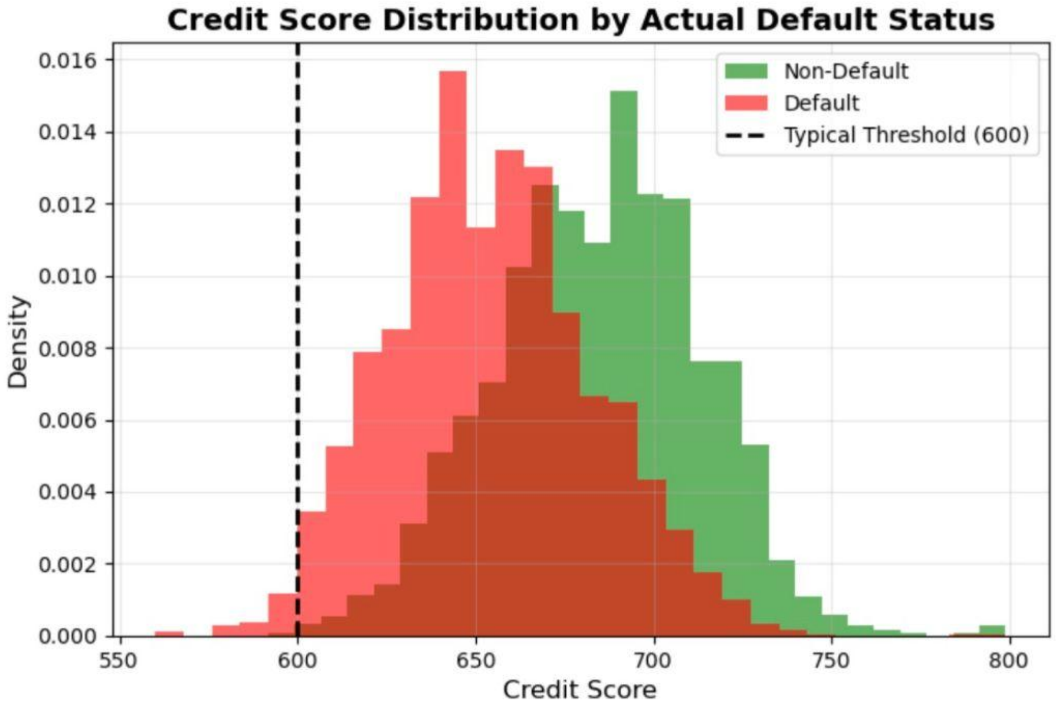
Threshold Analysis & Confusion Matrices

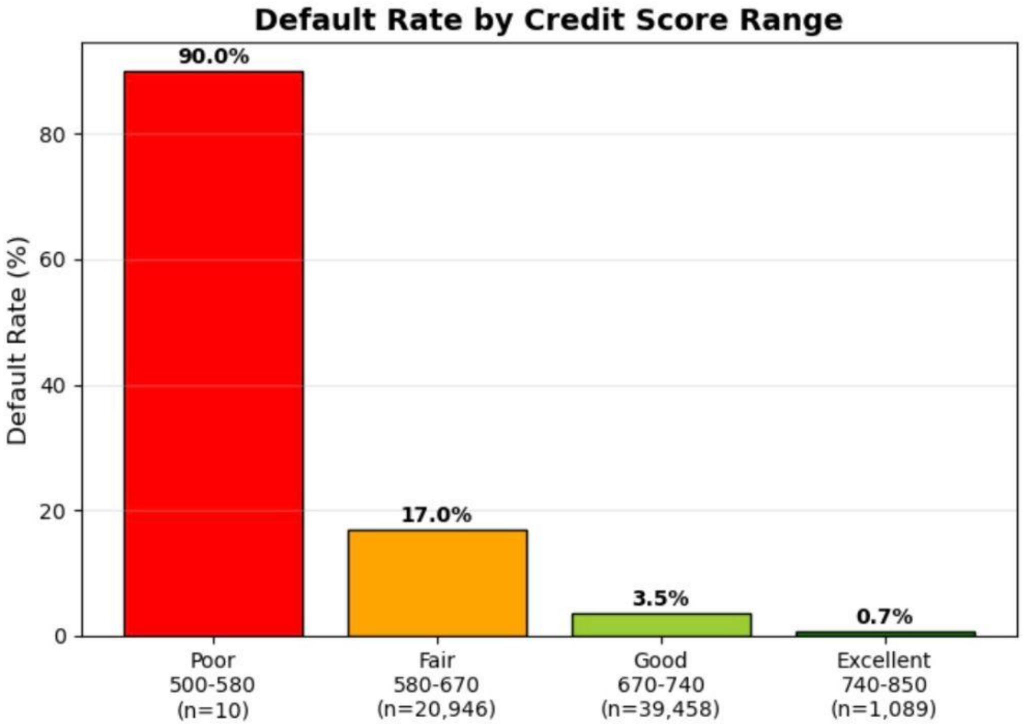


Key Insights:

- ❖ No single "best" threshold → depends on business strategy
- ❖ Lower threshold = more loans rejected but fewer defaults slip through
- ❖ Recommended: 0.15-0.25 range balances approval rate and risk
- ❖ Advantage of PD approach: Flexible threshold adjustment without retraining

Credit Score Application





Score = 600 + (20 / ln(2)) × ln((1-PD) / PD)

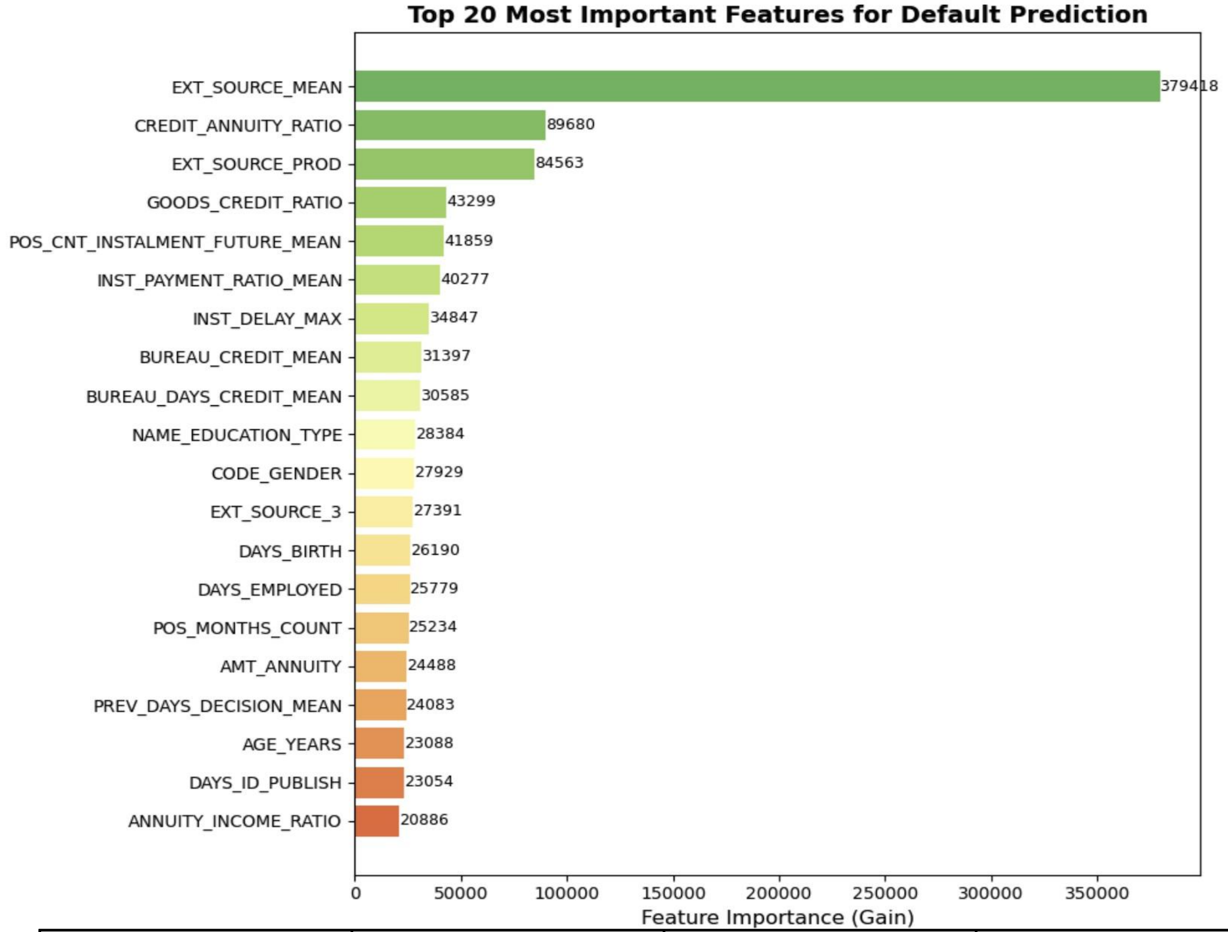
Where:

- BaseScore = 600 (odds 1:1)
- PDO = 20 (points to double odds)
- Higher PD → Lower Score

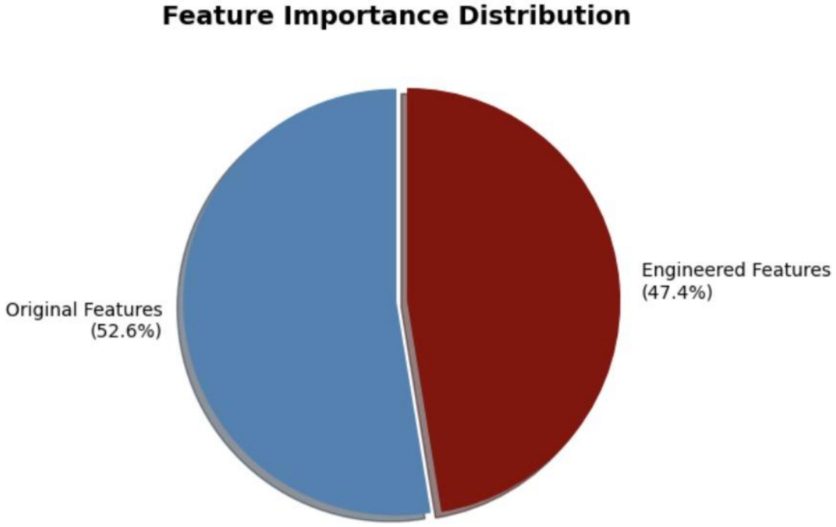
Approval Decisions:

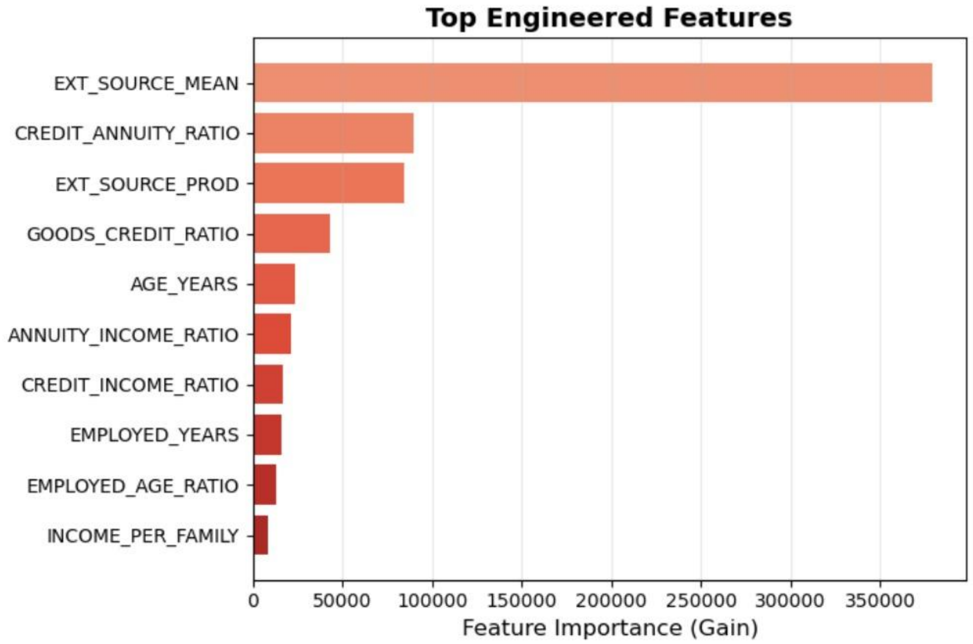
- ❖ **Score < 580:** Auto-reject or manual review
- ❖ **580-670:** Approve with higher rates/lower limits
- ❖ **670+:** Standard approval

Feature Importance Analysis



Metric	Without FE	With FE	Improvement
ROC-AUC	0.7739	0.7787	+0.62%
Brier Score	0.1785	0.1771	-0.81%
Avg Precision	0.2690	0.2719	+1.08%





03

Interactive Dashboard Demo