Rakamin Academy

HOME CREDIT

Approve This Loan?

Reducing Defaults & Boosting Revenue with Predictive Models

Presented by

Muhammad Rizdky Maulady

Project Based Internship - Data Scientist Home Credit Indonesia

TABLE OF CONTENT

Problem Research

1. Revenue
$653.41 B

Total Good Client		Revenue / Avg Loan
1306813	X	$500.000

2. Total Bad Client Debt
$61.67 B

Total Bad Client		Revenue / Avg Loan
123340	X	$500.000

3. Net Revenue
$591.74 B

Problem Statement

The current loan approval process lacks effective risk assessment, resulting in a high number of bad clients and significant revenue loss.

Goal

Reduce default rate through data-driven client assessment using machine learning.

Objective

Build and deploy a predictive model to identify high-risk clients before loan approval.

Business Metrics

- Default Rate Reduction (%)
- Increase in Net Revenue ($)
- Model Accuracy / Recall on Bad Clients (%)
- Cost Saved from Bad Loans ($)

"High loan default rate (8.6%) significantly reduces net revenue by over $61.67B annually."

Data Preparation

POS_CASH_Balance.csv
- No Duplicate
- Handling Missing values use Median
- Agg SK_DPD by SK_ID_PREV (mean)

credit_card_ balance.csv
- No Duplicate
- Handling Missing values use Median
- Agg(sum) 5 features by SK_ID_Prev

installments_ payments.csv
- No Duplicate
- Handling Missing values use drop
- Agg(sum) by ID

bureau_ balance.csv
- No Duplicate
- No Missing values
- Create new features

previous_ application.csv

bureau.csv
- No Duplicate
- Handling Missing value >50% drop, 5% s/d 50% median, <5% drop

application_train.csv

Main Dataset

1.430.155 rows x 105 feature

DATA PRE-PROCESSING

PRE-PROCESSING PIPELINE

01 Feature Engineering

- **Age Grouping**: Binned DAYS_BIRTH into age groups
- **Row Filtering**: Dropped NAME_FAMILY_STATUS = 'Unknown'
- **Missing Values**: Replaced 'XNA', 'XAP' with NaN
- **Value Correction**: Fixed CNT_FAM_MEMBERS anomalies (0.5 → 1, 4.5 → 5)

02 Data Splitting

Train 75 & Test 25, stratify=y

03 Feature Selection

- Removed high-missing-value features
- Selected variables with strong IV
- Dropped multicollinear features (corr ≥ 0.7)

04 Handling Outlier

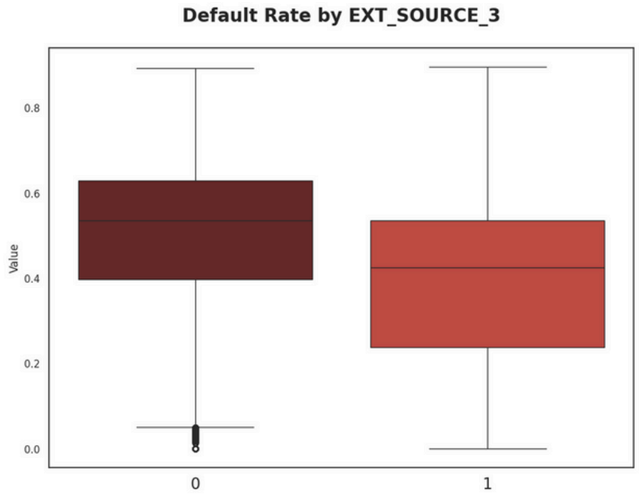
IQR Winsorization: Capped outliers at IQR boundaries

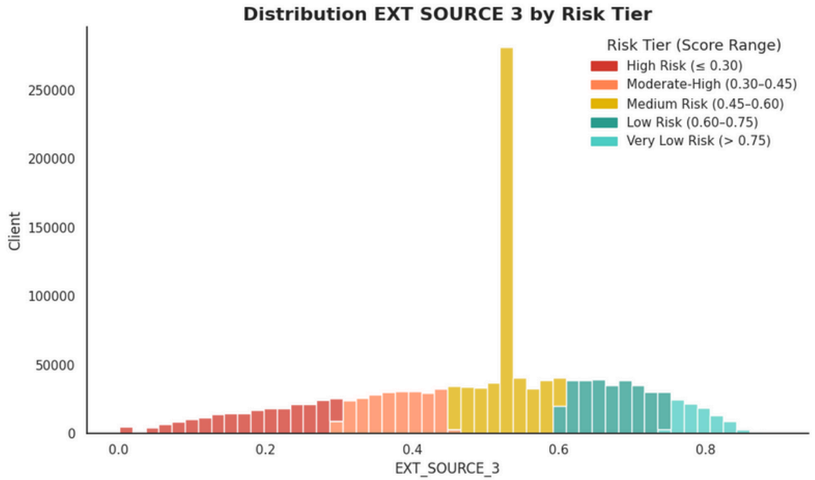
05 Feature Encoding & Scalling

Feature **Binning** & **WoE** Transformation

BUSINESS INSIGHT

Rakamin Academy HOME CREDIT





Insight:

- Clients with higher EXT_SOURCE_3 scores (above 0.45) tend to have lower default rates.
- The majority of clients fall into the medium to low risk category (score 0.45 – 0.75).
- This indicates a strong positive correlation between EXT_SOURCE_3 and client creditworthiness.

Recommendation:

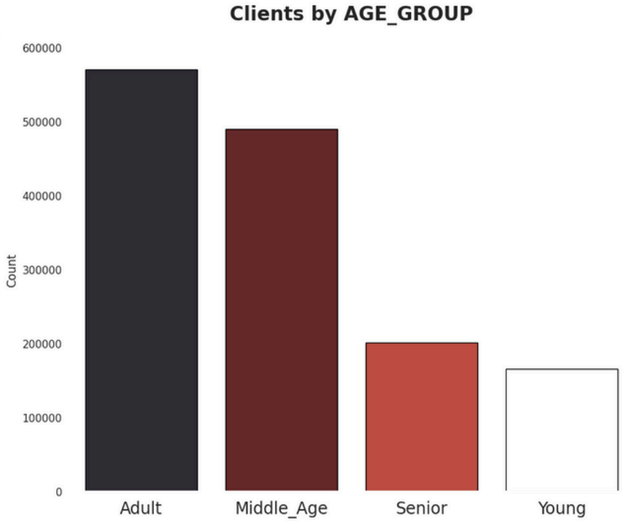
- Target segments with EXT_SOURCE_3 ≥ 0.45.
- Use AI-powered digital ads & lookalike audiences.
- Partner with relevant digital platforms (Socmed, E-commerce, etc.)

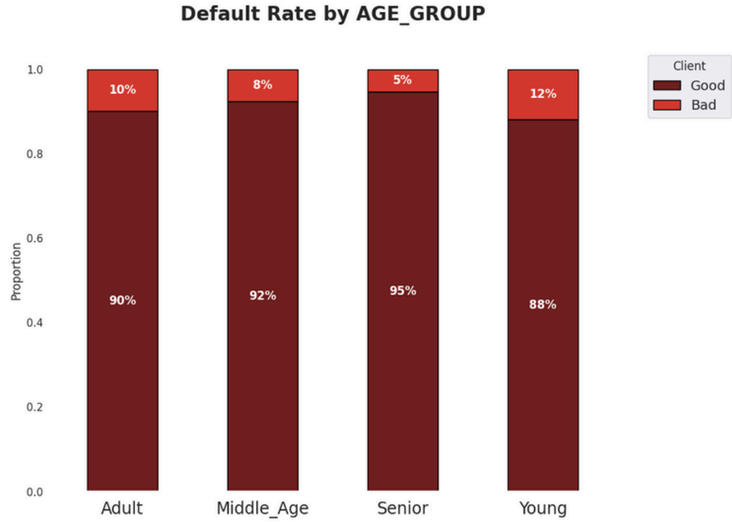
Insight:

- Most clients are in the Adult and Middle Age groups.
- Seniors show the lowest default rate (5%).
- Young clients have the highest credit risk (12% default rate).

Recommendation:

- Focus on retaining Adult and Middle Age clients.
- Use financial education & low-risk products for Young clients.
- Collaborate with platforms popular among younger demographics (social media, fintech apps, etc).



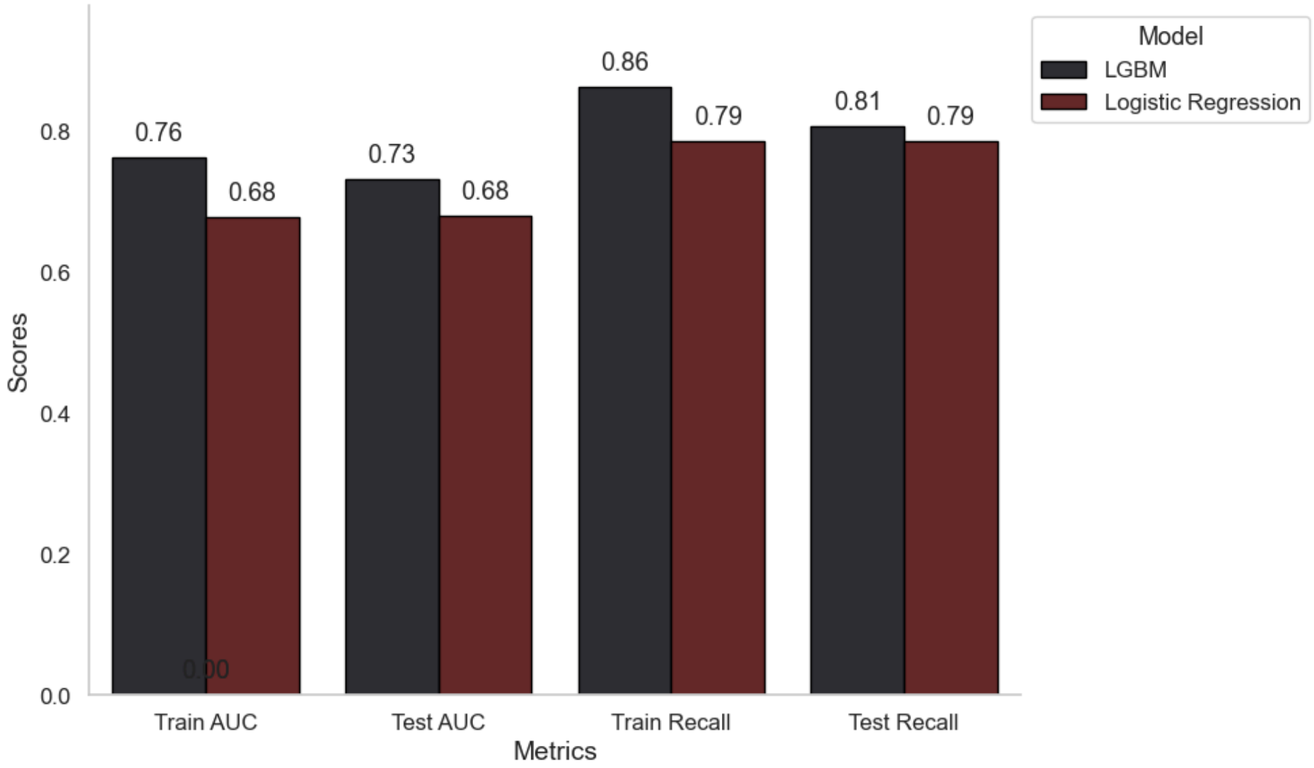


IMPLEMENTATION & EVALUATION

Metrics Evaluation : Recall

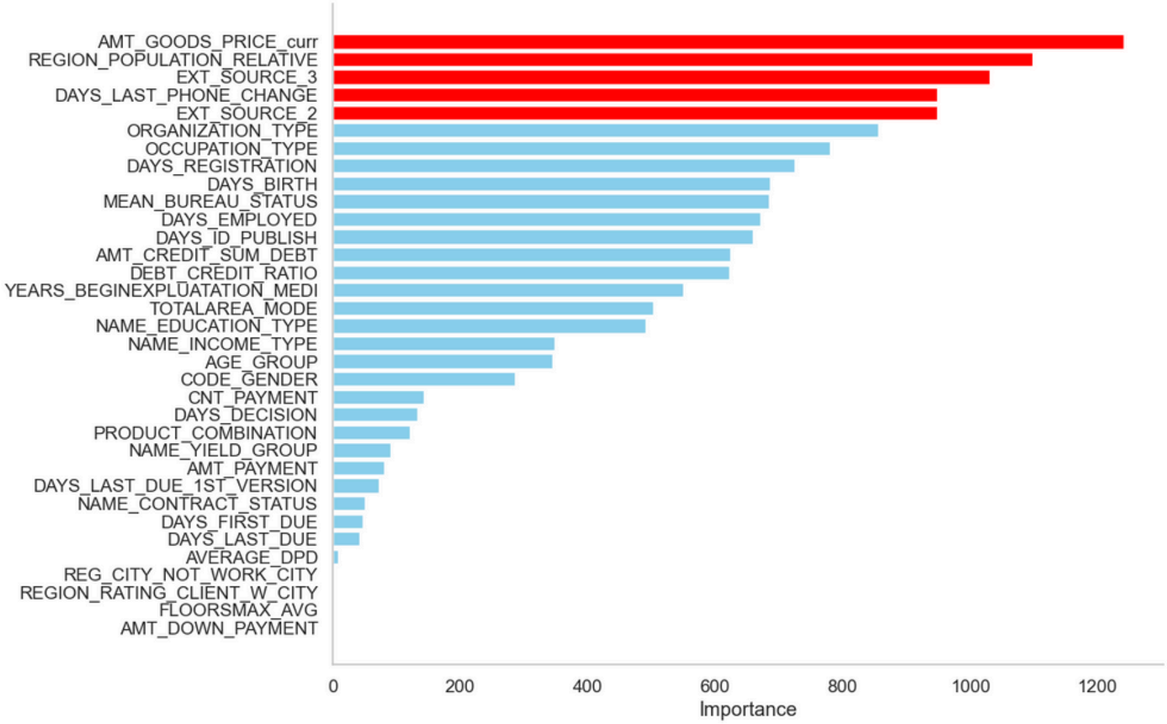
Recall measures the model's ability to detect high-risk clients (actual positives). A high recall score indicates better performance in capturing most default cases.

Model Comparison - Hyperparameter Tuning

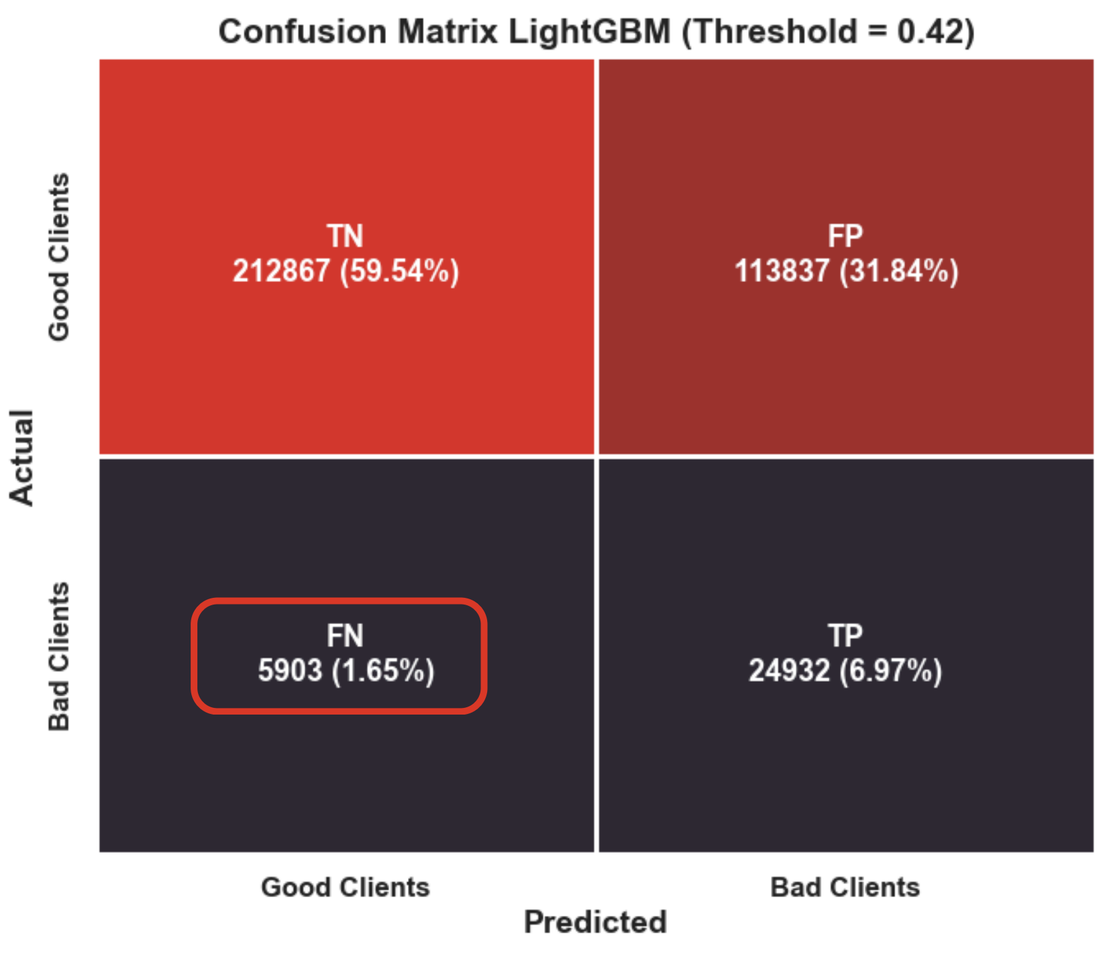


Feature Importances : LightGBM

Best Model LightGBM
class_weight='balanced'{'num_leaves': 150, 'max_depth': 30}

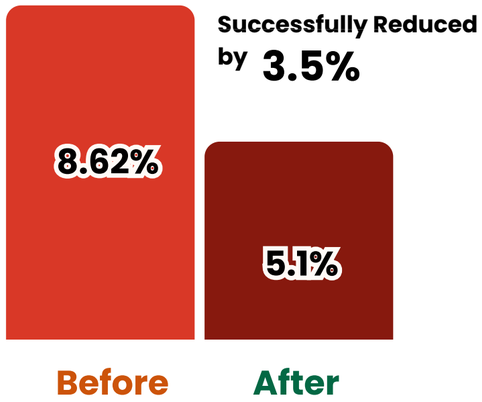


IMPLEMENTATION & EVALUATION



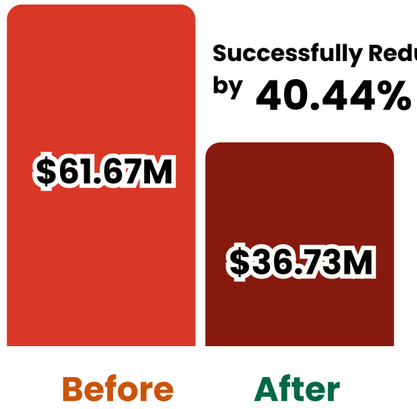
Impact Simulation

Default Rate



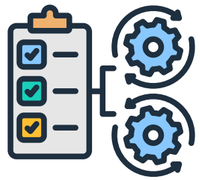
Increase in Net Revenue ($)

Cost Saved from Bad Loans ($)



BUSINESS RECOMMENDATION

Offer Credit Schemes Based on Risk Score & Product Value



Insight: Clients with EXT_SOURCE_3 ≥ 0.45 show significantly lower default rates.

Strategy:

- Leverage ML-driven credit scoring (EXT_SOURCE_3 & EXT_SOURCE_2) to personalize loan limits, interest rates, and tenors.
- Automate installment offers based on product price (AMT_GOODS_PRICE) and customer risk profile.
- Implement real-time loan approval with transparent interest rates tailored to individual risk scores.
- quality.

Target High-Population Areas with Localized Campaigns

Insight: Customers from densely populated regions (REGION_POPULATION_RELATIVE) tend to be more digitally active and lower-risk.

Strategy:

- Use geo-AI and population heatmaps to prioritize marketing and digital expansion areas.
- Launch hyper-local campaigns with customized content (e.g., regional languages, local influencers).
- Deploy digital financial literacy programs in high-potential but underbanked regions.

Monitor Digital Behavior Changes as Risk Signals

Insight: Customers from densely populated regions (REGION_POPULATION_RELATIVE) tend to be more digitally active and lower-risk.

Strategy:

- Implement early warning systems based on changes in digital identity (phone, email, device).
- Trigger verification or alerts when sudden digital behavior shifts are detected.
- Combine digital footprint data to strengthen fraud detection and identity validation models.

Thank You

"Insights are clear, today’s decisions shape tomorrow’s growth."

Github

Linkedin

Website