Loan Default Prediction

FROM RAW DATA TO ACTIONABLE RISK SCORES

- Goal: predict probability of default for each loan
- Help lender reduce losses and manage risk proactively
- Evaluation metric: AUC on highly imbalanced data (5.1% defaults)

Renaissance Technologies

Team:
Dilshod Ikromov
Shahriyor Akramov
Dilshod Azamjonov
Mukaddas Khusniddinova

DATA & CHALLENGES

- Source: historical loan and client information (6 different sources)
- Target: default vs non-default (5.1% default rate, ~18.6:1 ratio)
- Main challenges:
  - Class imbalance
  - Noisy / heterogeneous features
  - Need for robust, not overfitted model

OUR APPROACH – END-TO-END PIPELINE

Step 1

Data consolidation & cleaning

Merge raw tables, standardize IDs and types, handle missing values/outliers, create a single modeling dataset.

Step 2

Exploratory analysis & target checks

Analyze distributions, class imbalance and correlations, validate target definition, check for data leakage and basic sanity.

Step 3

Feature engineering & encoding (111 features, 17 engineered)

Build financial ratios, credit-behaviour and geo features; apply one-hot + 5-fold target encoding for high-cardinality categoricals.

Step 4

Modeling & validation

Train logistic baseline → XGBoost / LightGBM with class-imbalance handling; tune hyperparameters with 5-fold CV; select a stacking ensemble as final model.

Step 5

Threshold tuning & probability calibration

Optimize decision threshold, calibrate predicted PDs to portfolio default rate, and export a production-ready model + score file.

Renaissance Technologies

MODELS & QUANTITATIVE RESULTS

Model	AUC	Precision	Recall	F1-Score	Overfit Gap
Logistic Regression	0.7927	0.4157	0.2862	0.3390	0.1567
XGBoost	0.7998	0.4184	0.2988	0.3485	0.1792
LightGBM Baseline	0.7941	0.4212	0.2819	0.3382	0.1567
LightGBM Optimized	0.8044	0.4212	0.2884	0.3419	0.0278
Stacking Ensemble	0.8074	0.4202	0.2884	0.3411	0.0264

WHY IT WORKS & WHAT THE MODEL LEARNED?

0.8074 AUC

robust risk separation

Our optimized **LightGBM + stacking ensemble** achieves 0.8074 ROC-AUC on an imbalanced dataset with **5.1%** defaults. This shows the model can reliably rank high-risk vs low-risk borrowers, even under heavy noise and skewed distributions.

118 engineered features

robust risk separation

We created a structured feature space with 118 variables, combining financial ratios, credit behavior metrics, geographic risk patterns, and smoothed target-encoded categoricals. This enables the model to capture non-linear relationships that simple baselines miss.

44% reduction in overfitting

generalization through tuning

Hyperparameter optimization (30-iteration randomized search over 5-fold CV) reduced the LGBM train–test AUC gap from 0.1567 → 0.0445. This demonstrates that our model does not memorize noise and will generalize to unseen borrowers.

Learned behavioral patterns

interpretable risk drivers

The model consistently identifies:

- Low income-to-loan ratio → high probability of default
- High debt burden and deteriorating credit age
- Certain employment-type categories with elevated risk
- Geographic clusters correlated with repayment difficulty

Renaissance Technologies

Thank you!

Renaissance Technologies