Credit Risk Assessment for Financial Institutions

Anahad Singh, Manavi Nakra, Tanmay Mohan

PROBLEM STATEMENT

LENDER: “SHALL I PROVIDE A LOAN?”

While loans help people access money when they need it, lenders must decide whether a borrower will be able to repay the loan. Predicting this repayment ability is a major challenge in the financial industry.

STATUS QUO

Banks predict default by looking at a static snapshot of your money today (e.g., total debt, income). They completely ignore your behavior over time.

Banks get blindsided when seemingly "safe" borrowers suddenly run out of cash. Meanwhile, they unfairly reject responsible people just because they have a short credit history.

THE PROBLEM

A LOAN DEAFULT occurs when a borrower fails to repay the loan according to the agreed repayment schedule.

How will Financial Institutions predict loan default risk?

LITERATURE SURVEY

Literature Review

Credit default prediction is tricky because for an applicant their risk is spread across application details, bureau records, previous loans, credit cards, installments, and monthly repayment behavior.

Efficient Commercial Bank Customer Credit Risk Assessment Based on LightGBM and Feature Engineering

Sun Yanjie *
School of economics and management
University of Electronic Science and Technology of China
Chengdu, China
1789660136@qq.com

Gong Zhike*
School of economics and management
University of Electronic Science and Technology of China
Chengdu, China
2020150501004@std.uestc.edu.cn

Shi Quan
Information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
563832834@qq.com

Chen Lin†
School of economics and management
University of Electronic Science and Technology of China
Chengdu, China
chenlin2@uestc.edu.cn

Sun et al. use LightGBM with feature engineering. They handle missing values, categorical encoding, imbalance, and create a few simple ratio features.

Sun, Y., Gong, Z., Shi, Q., & Chen, L. (2024). Efficient commercial bank customer credit risk assessment based on LightGBM and feature engineering. IEEE Access. https://ieeexplore.ieee.org

Relational Graph Modeling for Credit Default Prediction: Heterogeneous GNNs and Hybrid Ensemble Learning

Yvonne Yang
University of Illinois Urbana-Champaign
yuweny4@illinois.edu

Eranki Vasistha
University of Illinois Urbana-Champaign
veranki2@illinois.edu

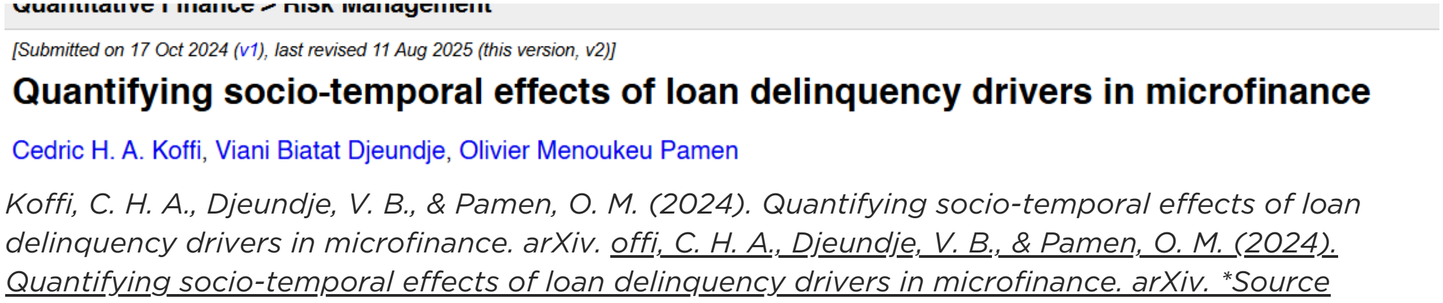
Yang and Vasistha go in another direction. They treat the same dataset as a graph and use GNN embeddings with LightGBM

Yang, Y., & Vasistha, E. (2026). Relational graph modeling for credit default prediction: Heterogeneous GNNs and hybrid ensemble learning. https://arxiv.org/abs/2501

THE GAP

Economics literature often points to what things cause an applicant to default already, which ML studies miss, most economics studies point to questions like -

Can they afford this loan?
Have they started paying late recently?
Is their delinquency getting worse?
Are they using too much credit?
Have they been getting refused repeatedly?
Are they showing signs of stress before default?

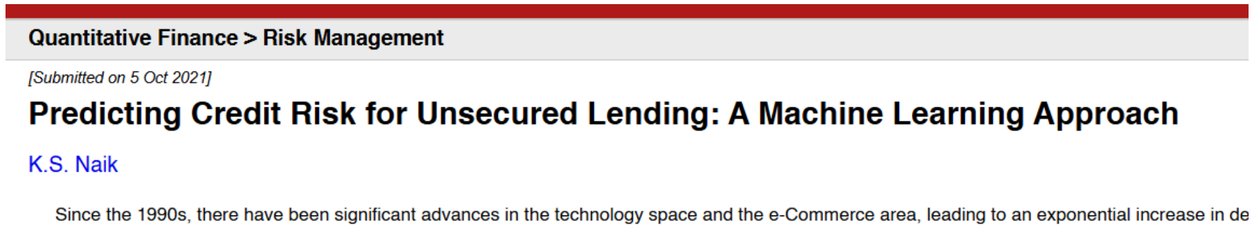

Quantitative Finance > Risk Management

[Submitted on 17 Oct 2024 (v1), last revised 11 Aug 2025 (this version, v2)]

Quantifying socio-temporal effects of loan delinquency drivers in microfinance

Cedric H. A. Koffi, Viani Biatat Djeundje, Olivier Menoukeu Pamen

*Koffi, C. H. A., Djeundje, V. B., & Pamen, O. M. (2024). Quantifying socio-temporal effects of loan delinquency drivers in microfinance. arXiv. offi, C. H. A., Djeundje, V. B., & Pamen, O. M. (2024). Quantifying socio-temporal effects of loan delinquency drivers in microfinance. arXiv. *Source*


Quantitative Finance > Risk Management

[Submitted on 5 Oct 2021]

Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach

K.S. Naik

Since the 1990s, there have been significant advances in the technology space and the e-Commerce area, leading to an exponential increase in de

*Naik, K. S. (2021). Predicting credit risk for unsecured lending: A machine learning approach. arXiv. offi, C. H. A., Djeundje, V. B., & Pamen, O. M. (2024). Quantifying socio-temporal effects of loan delinquency drivers in microfinance. arXiv. *Source*

Prior Home Credit work already uses feature engineering.

Credit-risk ideas are usually not shown as explicit time-aware borrower-behavior features.
So focus became recent stress, worsening delinquency, debt pressure, utilization, refusal velocity, and repayment irregularity.

DATASET

Home Credit Default Risk
Kaggle Competition in 2018

Description

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

WHY

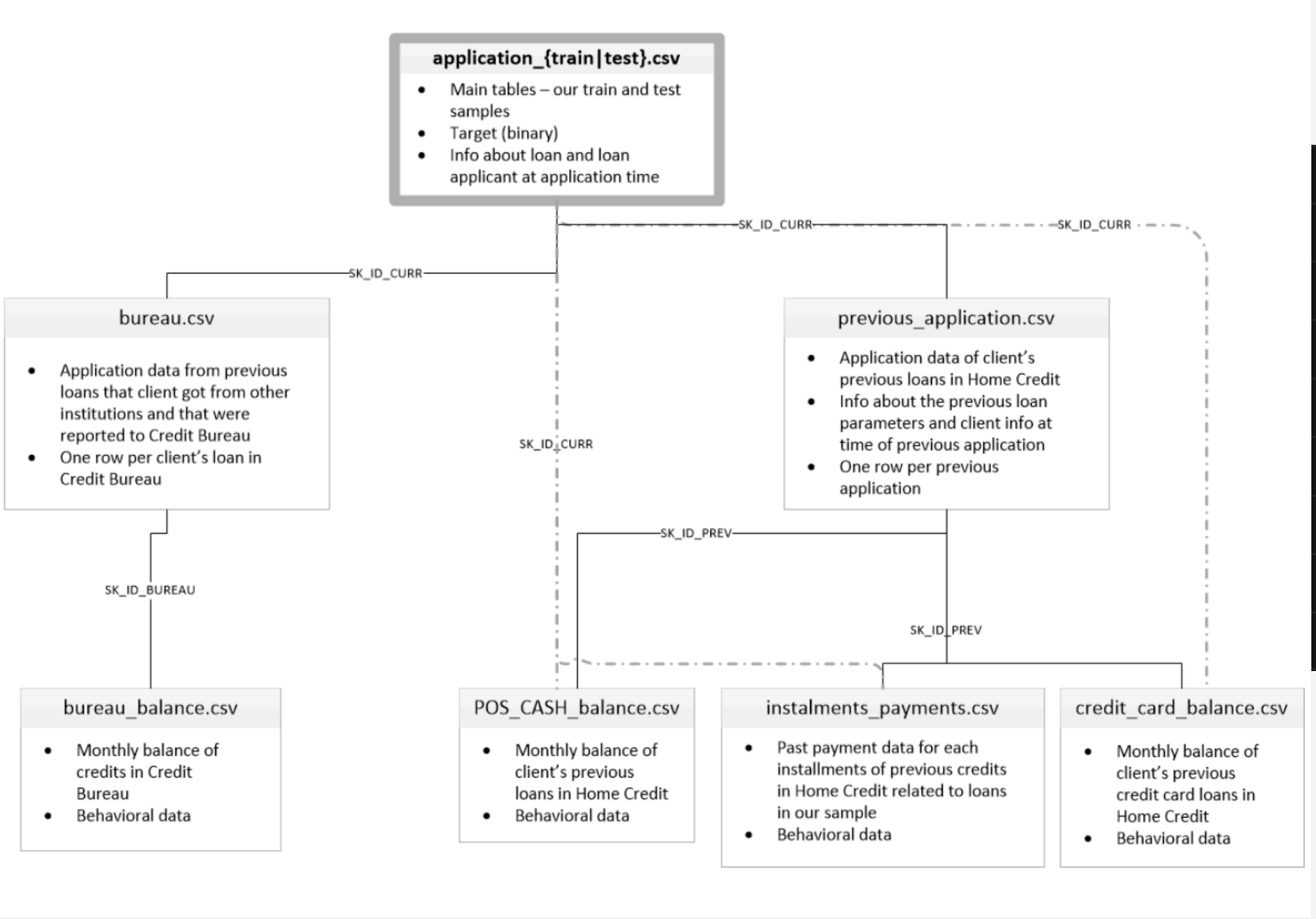
It contains large-scale financial data suitable for machine learning models. It includes multiple related tables, allowing rich feature engineering.

Real financial datasets are difficult to curate due to privacy, legal, and ethical restrictions, so using a well documented dataset released for research is appropriate.

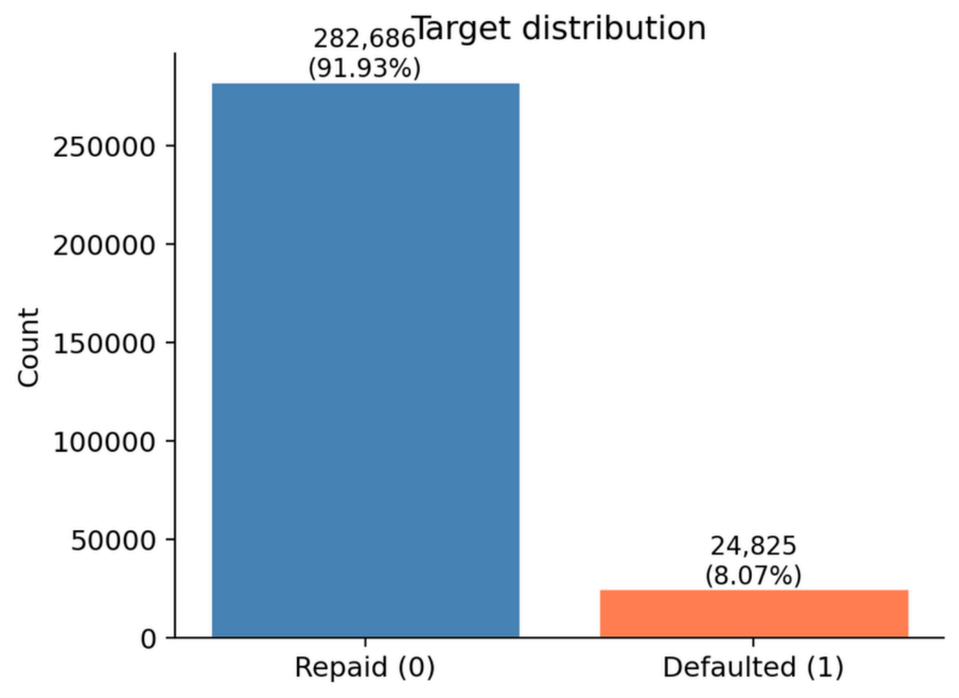
ETHICAL CONSIDERATIONS

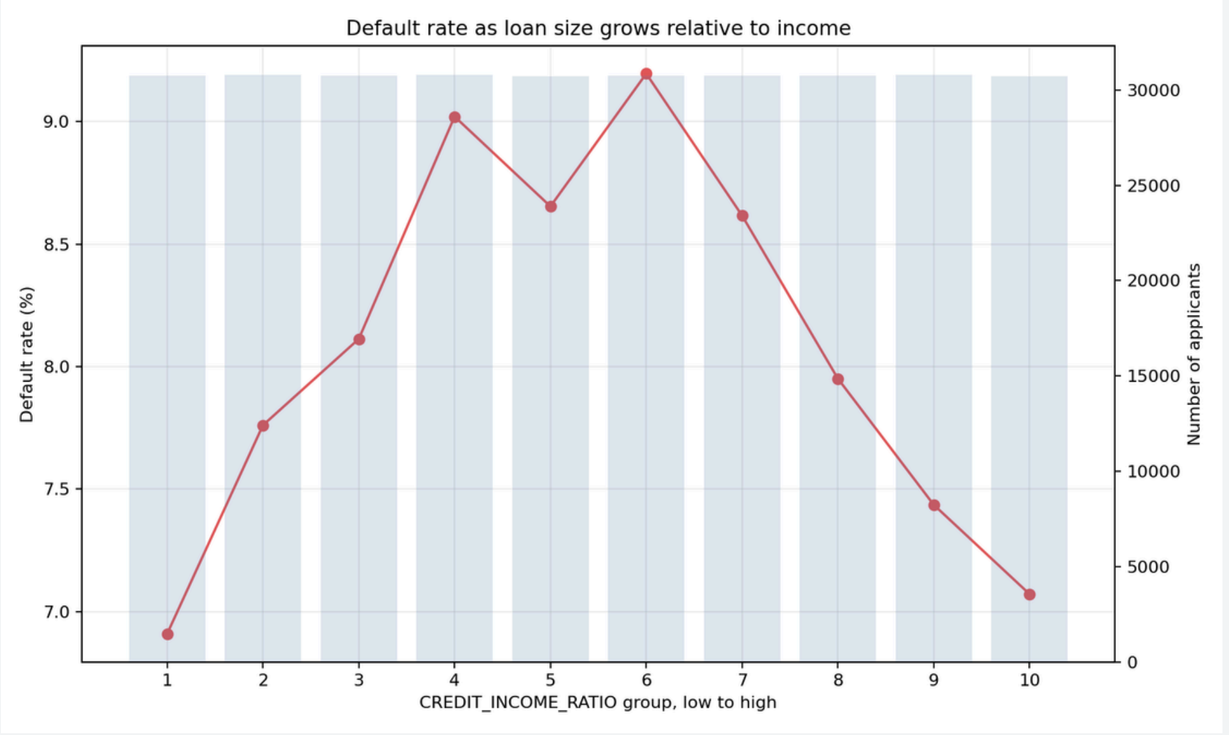
- Dataset was anonymized to protect borrower privacy
- Personal identifiers were removed
- Shared publicly only for research and educational purposes

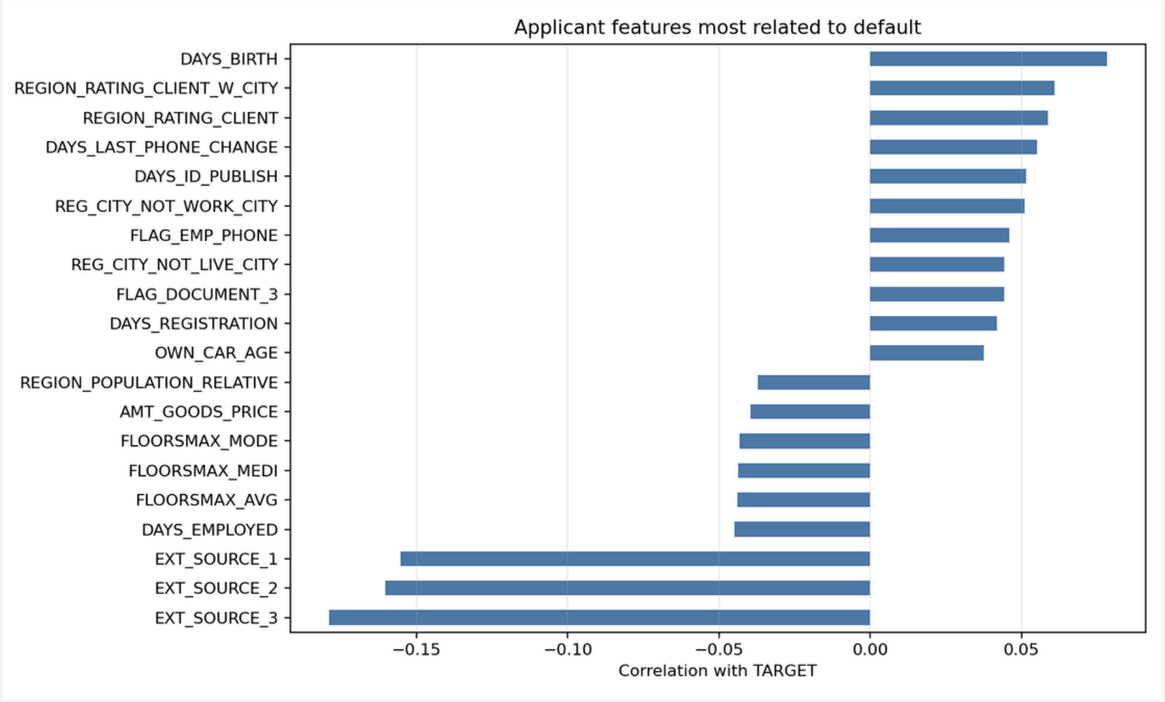
DATASET INFO

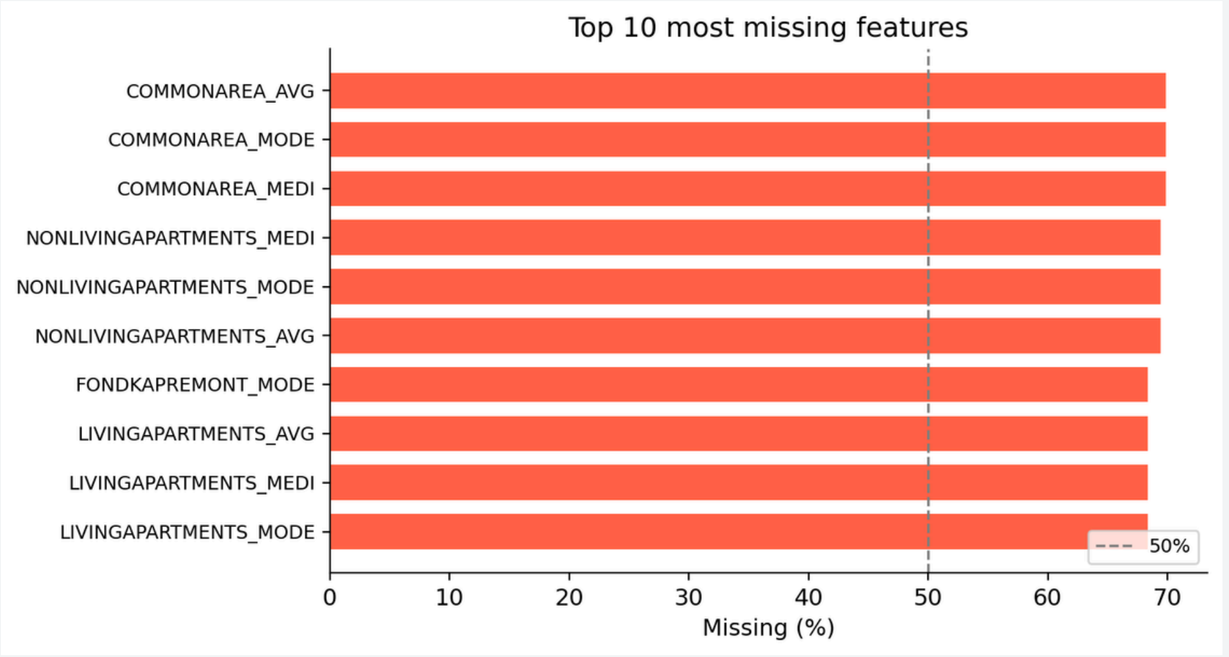


File Name	Rows	Columns
application_train.csv	307,511	122
application_test.csv	48,744	121
bureau.csv	1,716,428	17
bureau_balance.csv	27,299,925	3
credit_card_balance.csv	3,840,312	23
installments_payments.csv	13,605,401	8
POS_CASH_balance.csv	10,001,358	8
previous_application.csv	1,670,214	37


Target distribution
282,686
(91.93%)
24,825
(8.07%)
Count
250000
200000
150000
100000
50000
0
Repaid (0)
Defaulted (1)


Default rate as loan size grows relative to income
Default rate (%)
Number of applicants
CREDIT_INCOME_RATIO group, low to high


Applicant features most related to default
DAYS_BIRTH
REGION_RATING_CLIENT_W_CITY
REGION_RATING_CLIENT
DAYS_LAST_PHONE_CHANGE
DAYS_ID_PUBLISH
REG_CITY_NOT_WORK_CITY
FLAG_EMP_PHONE
REG_CITY_NOT_LIVE_CITY
FLAG_DOCUMENT_3
DAYS_REGISTRATION
OWN_CAR_AGE
REGION_POPULATION_RELATIVE
AMT_GOODS_PRICE
FLOORSMAX_MODE
FLOORSMAX_MEDI
FLOORSMAX_AVG
DAYS_EMPLOYED
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
Correlation with TARGET


Top 10 most missing features
COMMONAREA_AVG
COMMONAREA_MODE
COMMONAREA_MEDI
NONLIVINGAPARTMENTS_MEDI
NONLIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_AVG
FONDKAPREMONT_MODE
LIVINGAPARTMENTS_AVG
LIVINGAPARTMENTS_MEDI
LIVINGAPARTMENTS_MODE
50%
Missing (%)

FEATURES PREPROCESSING

Sign normalisation: all DAYS columns stored negative → converted to positive

Outlier cap: car age clipped at 99th percentile

Log transform: income (right-skewed)

Missing values: trees handle natively; median imputation for KNN branch only

Categorical encoding: fold-safe target encoding (LGBM/XGB), native cats (CatBoost)

Zero-importance drops: ~30 features removed after iterative runs

EXAMPLES:

- DAYS_EMPLOYED = 365,243 → NaN (sentinel value)
- All DAYS_* columns → absolute value (stored as negative in raw data, e.g. -1000 = 1000 days ago)
- CatBoost: raw categoricals kept as-is; missing values filled with string "MISSING"

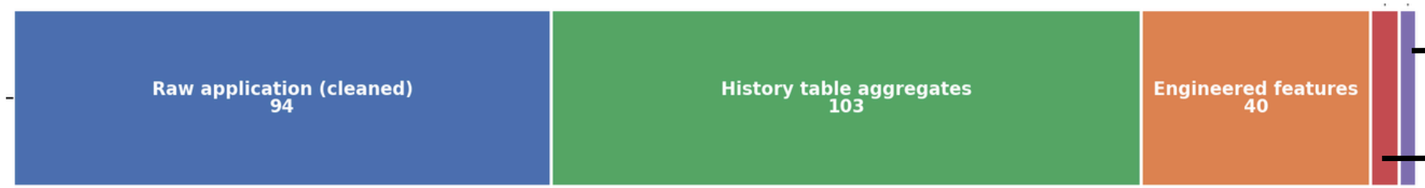
Zero importance feature dropping

```python
COLS_TO_DROP: list[str] = [
    # Run 1 — zero-importance across all 3 folds
    "FLAG_MOBIL", "FLAG_EMP_PHONE", "FLAG_CONT_MOBILE",
    "FONDKAPREMONT_MODE", "NAME_INCOME_TYPE",
    "FLAG_DOCUMENT_2", "FLAG_DOCUMENT_4", "FLAG_DOCUMENT_5",
    "FLAG_DOCUMENT_7", "FLAG_DOCUMENT_9", "FLAG_DOCUMENT_10",
    "FLAG_DOCUMENT_11", "FLAG_DOCUMENT_12", "FLAG_DOCUMENT_13",
    "FLAG_DOCUMENT_14", "FLAG_DOCUMENT_15", "FLAG_DOCUMENT_16",
    "FLAG_DOCUMENT_17", "FLAG_DOCUMENT_19", "FLAG_DOCUMENT_20",
    # Run 2 — zero-importance across all 3 folds
    "NAME_HOUSING_TYPE", "EMERGENCYSTATE_MODE", "FLAG_DOCUMENT_21",
    "AMT_REQ_CREDIT_BUREAU_HOUR", "AMT_REQ_CREDIT_BUREAU_DAY",
    "DOCUMENT_COUNT",
    # Run 3 — zero-importance across all 3 folds
    "HOUSETYPE_MODE", "WALLSMATERIAL_MODE", "DAYS_EMPLOYED_ANOM",
    # Run 4 — zero-importance across all 3 folds
    "REG_REGION_NOT_LIVE_REGION",
    # Run 12 — zero-importance across all 3 folds
    "FLAG_EMAIL",
    # Zero importance after meta-features
    "BUREAU_BB_STATUS_3_COUNT_SUM", "BUREAU_BB_STATUS_4_COUNT_SUM", "BUREAU_BB_STATUS_5_COUNT_SUM",
    # Zero importance in accepted run_id=bcf3118873d54371acb775a8d0d5e188
    "BUREAU_PROLONG_SUM", "BUREAU_BB_SEVERE_STATUS_SUM",
    # Zero importance in run_id=c599f7fbd26746bbab8eca3355bdcb9a
    "FLAG_DOCUMENT_6",
]
```

From 7 raw tables → 245 features per applicant

ANNUITY_INCOME_RATIO	AMT_ANNUITY / AMT_INCOME_TOTAL	Can income support yearly repayment?
CREDIT_INCOME_RATIO	AMT_CREDIT / AMT_INCOME_TOTAL	How large is the loan relative to income?
DAYS_LATE	max(DAYS_ENTRY_PAYMENT - DAYS_INSTALMENT, 0)	Was the borrower late on installments?
PAYMENT_RATIO	AMT_PAYMENT / AMT_INSTALMENT	Did they fully pay what was due?
INST_EWMA_LATE	weighted avg of late payments, higher weight to recent ones	Recent repayment stress
POS_DPD_ACCELERATION	recent max DPD – older max DPD	Is delinquency getting worse?
CC_UTILIZATION_RATIO	AMT_BALANCE / AMT_CREDIT_LIMIT_ACTUAL	Credit-card pressure

TOTAL_DTI	(ACTIVE_CREDIT_SUM + AMT_CREDIT) / (AMT_INCOME_TOTAL + 1)
TIME_DECAYED_DTI	(ACTIVE_CREDIT_SUM_DECAYED + AMT_CREDIT) / (AMT_INCOME_TOTAL + 1)
INST_LATE_FRAC	share of rows where DAYS_ENTRY_PAYMENT > DAYS_INSTALMENT
INST_PAYMENT_RATIO_MEAN	average AMT_PAYMENT / (AMT_INSTALMENT + ε)
CC_UTIL_MEAN_ALL	average balance / credit limit (where limit > 0)
PREV_REFUSED_FRAC	fraction of past applications with status "Refused"
PREV_2YR_REFUSED_VELOCITY	refused count in last 2 years ÷ apps in last 2 years



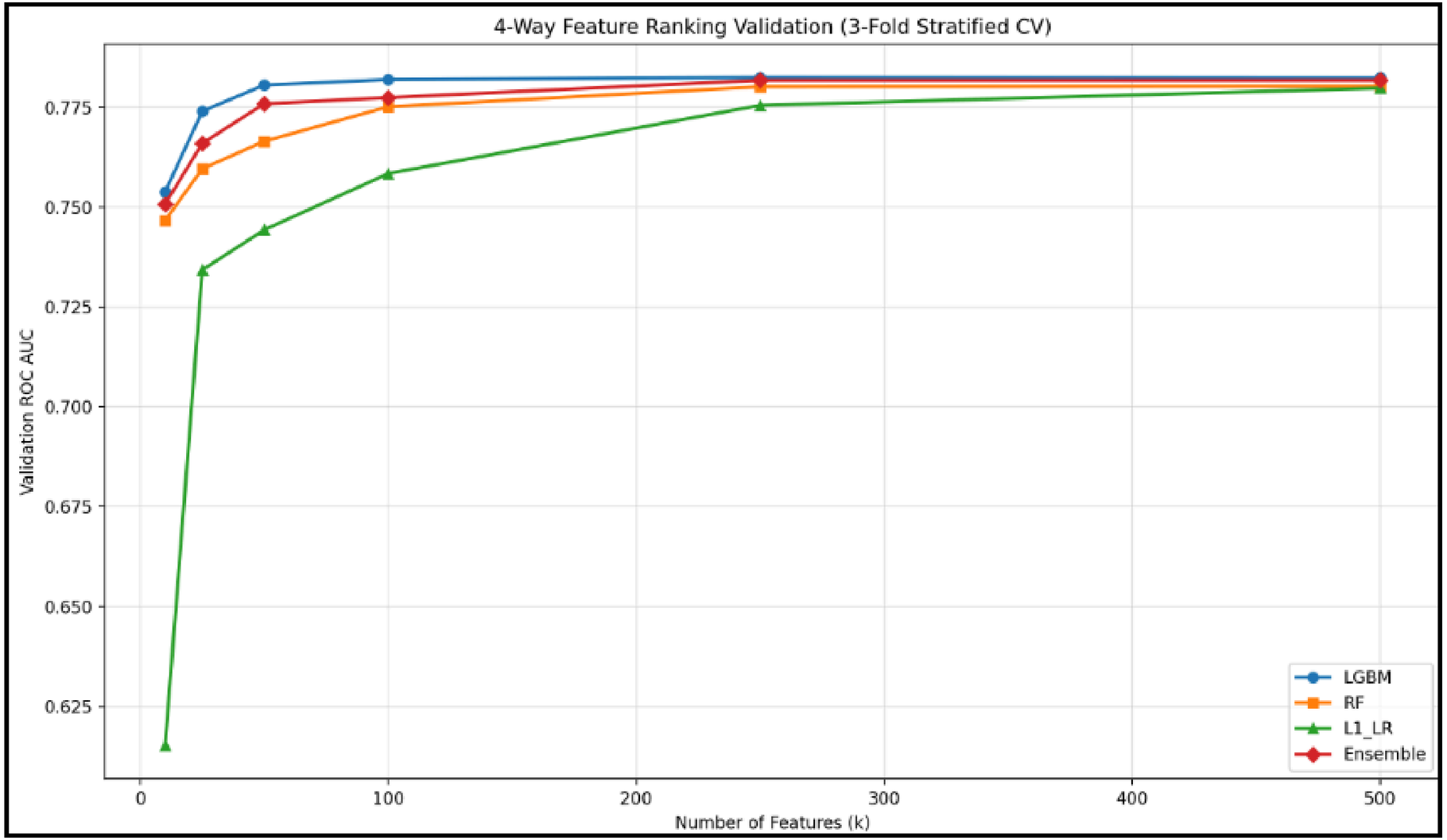
ML METHODOLOGY

WHICH MODEL?

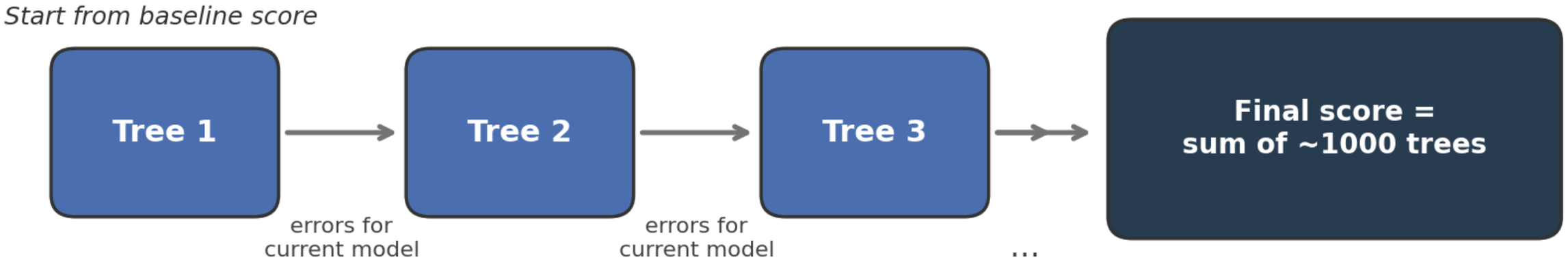
Model	Problem for this data	Verdict
Neural network	Needs feature scaling, doesn't handle missings natively, needs spatial/sequential structure	Not suitable
Logistic regression	Assumes each feature affects default linearly and independently — income × loan amount is not linear	Too simple
Gradient boosted trees	Handles missing values by learning which branch to take, captures non-linear interactions, dominates tabular benchmarks	**Our choice**

For mixed-type, missing-heavy tabular data, boosted trees are the boring-but-correct baseline.

Comparison between LGBM , RF and Logistic Regression for hypothesis testing

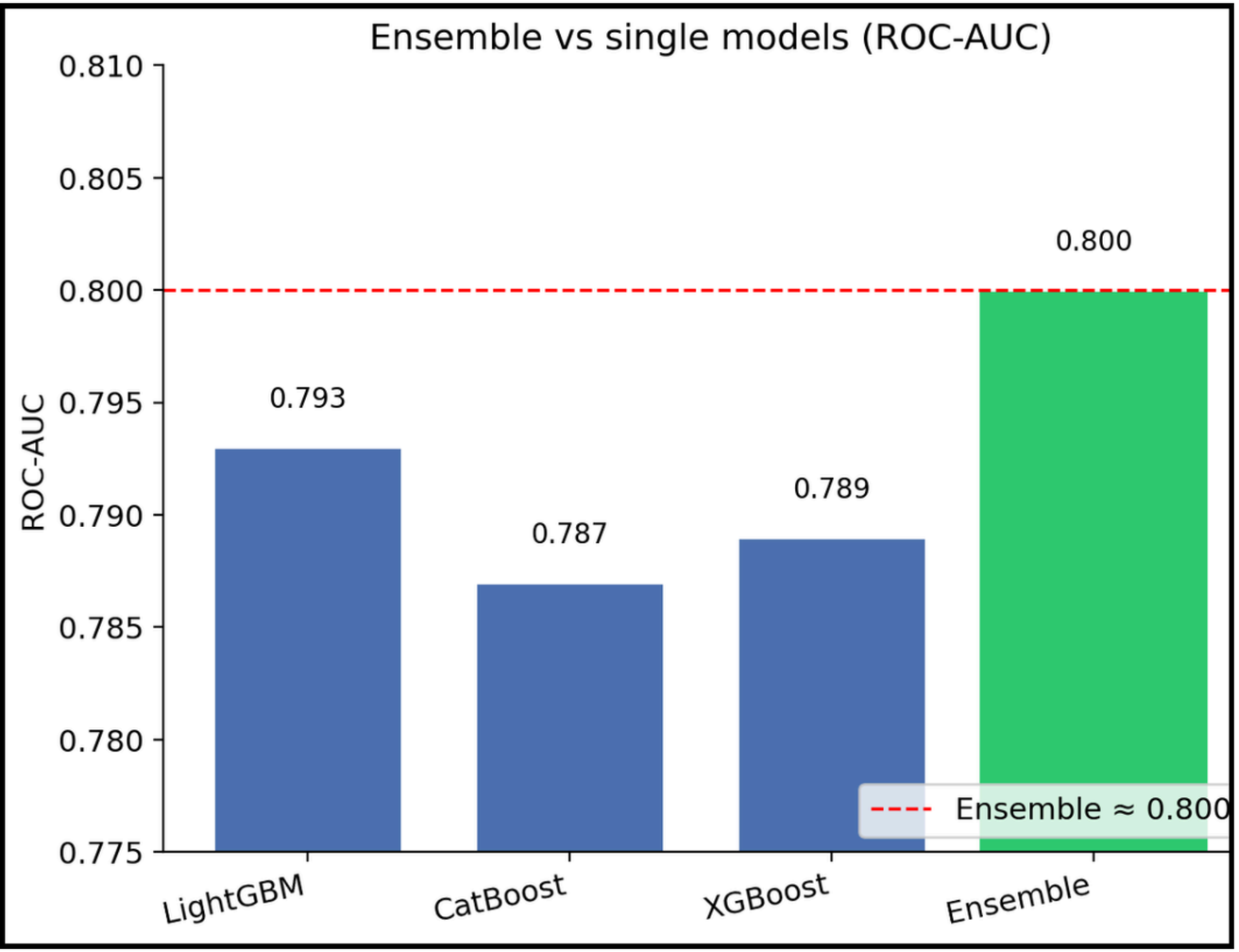


Gradient boosting — 1000 trees, each fixing the last



- **Each tree fits the gradient of the loss — the direction the current model is still wrong**
- **1000 rounds, learning rate 0.032 (tuned via Optuna) — small steps prevent any single tree from overfitting**
- **Final score = sum of all 1000 tree outputs on the log-odds scale**

LightGBM 65% · CatBoost 15% · XGBoost 20%



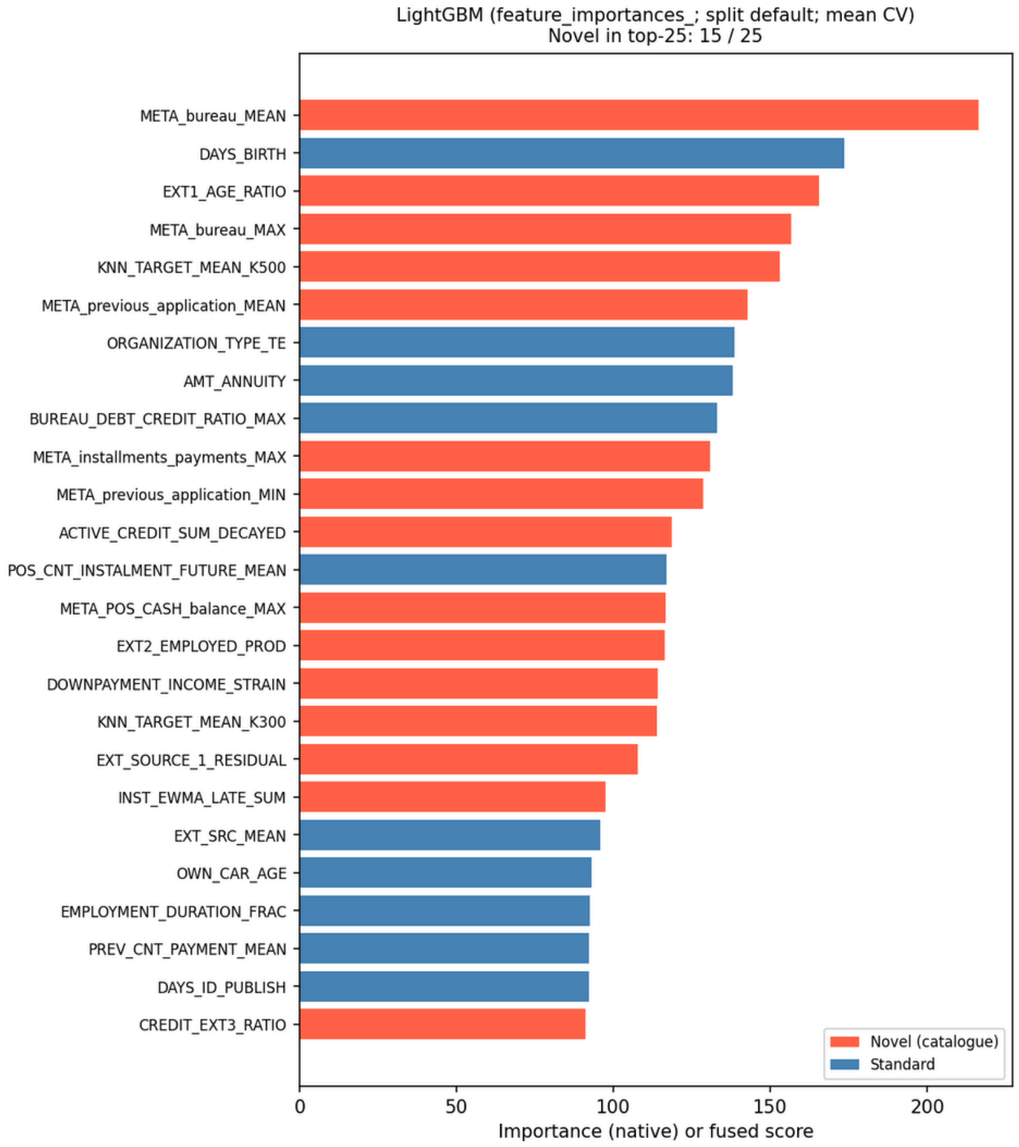
- **LightGBM — leaf-wise growth, fastest, highest individual AUC**
- **CatBoost — handles categoricals with ordered target statistics internally (no encoding needed)**
- **XGBoost — level-wise growth; structural diversity means errors don't perfectly overlap**

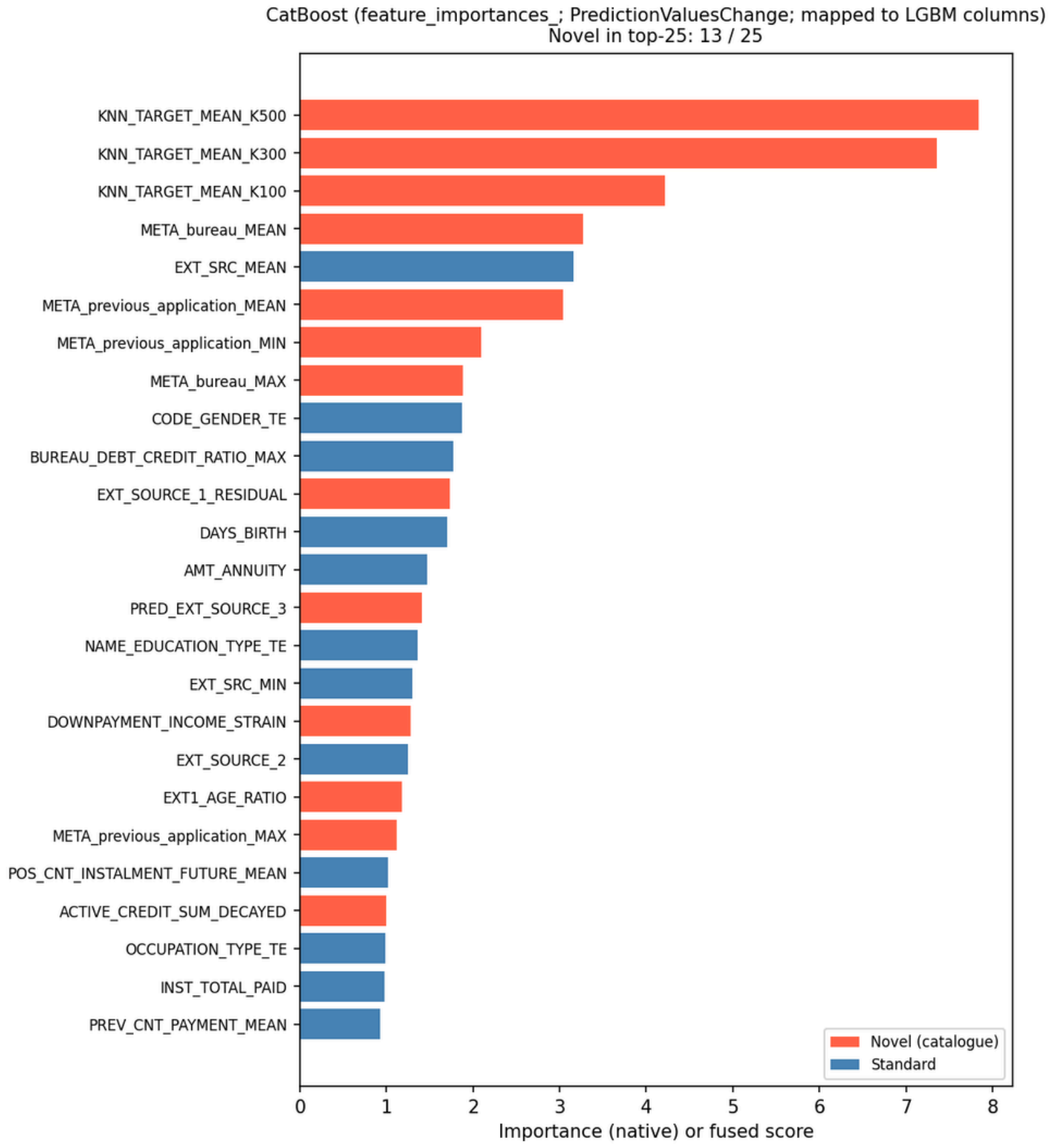
CHALLENGES

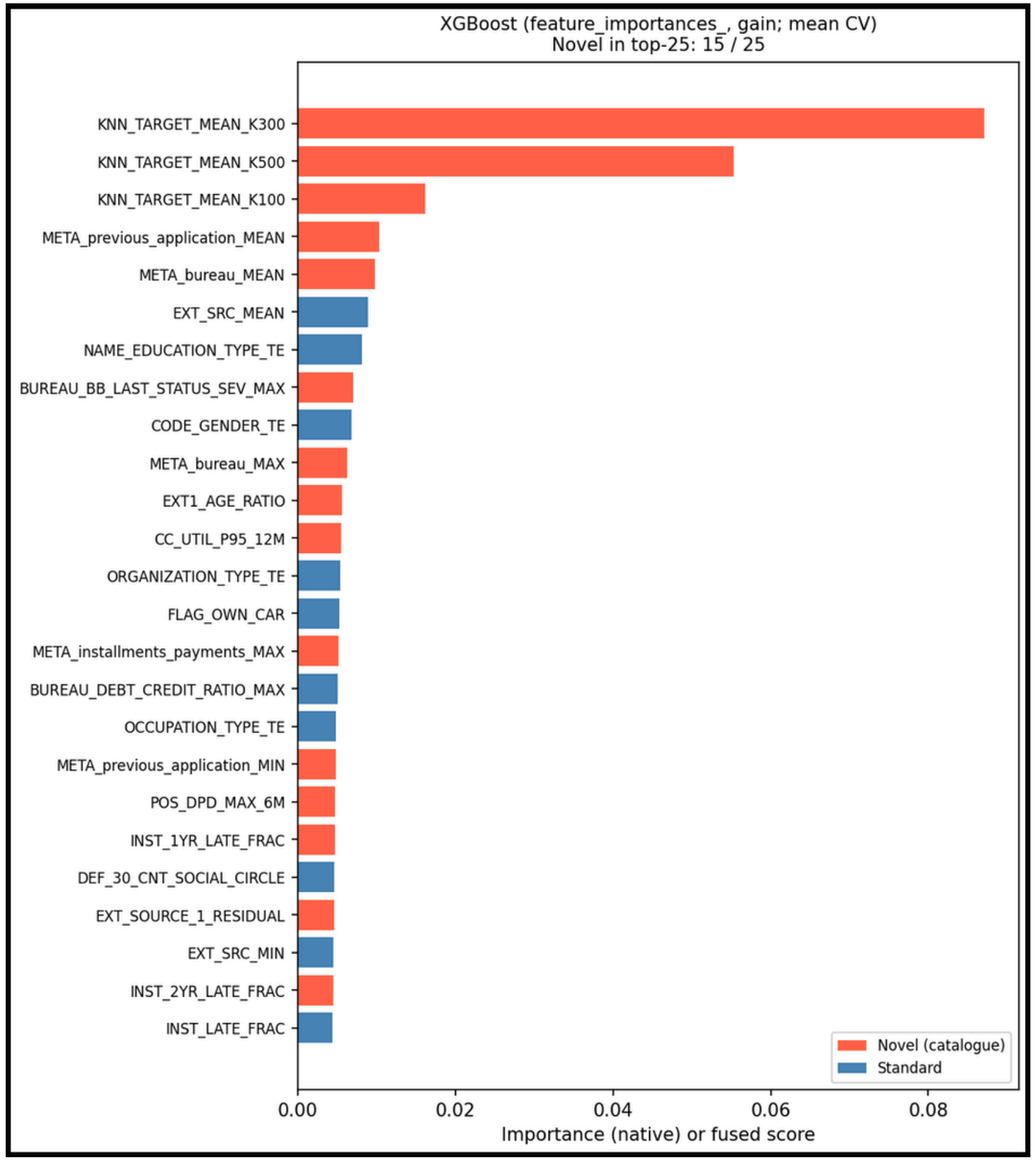
- Large volume data contributed to hardware issues - RAM overflow, GPU testing
- Preventing information leakage across three feature types
- Misleading values carry meaning in a highly sensitive credit risk dataset

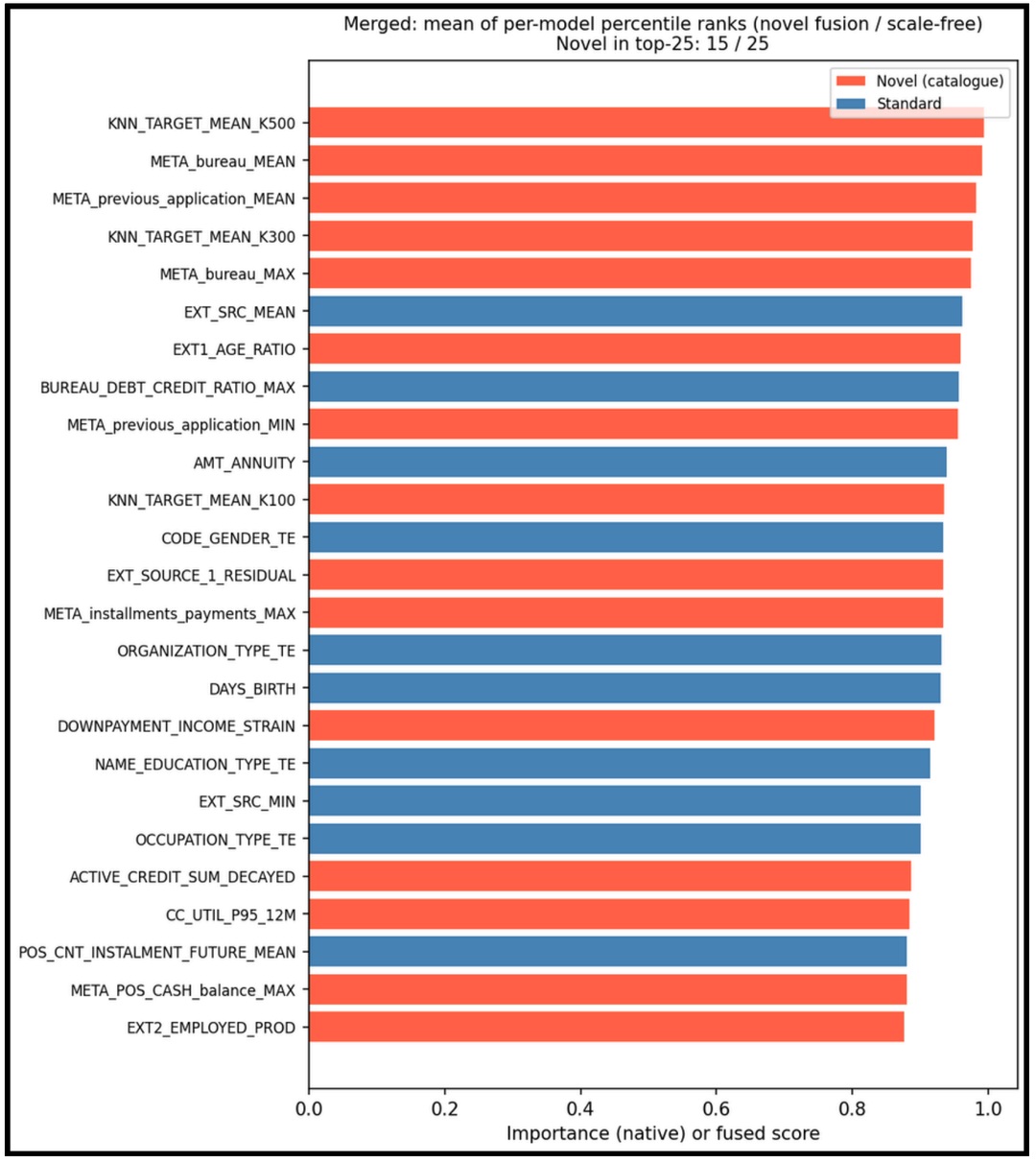
PERFORMANCE METRICS

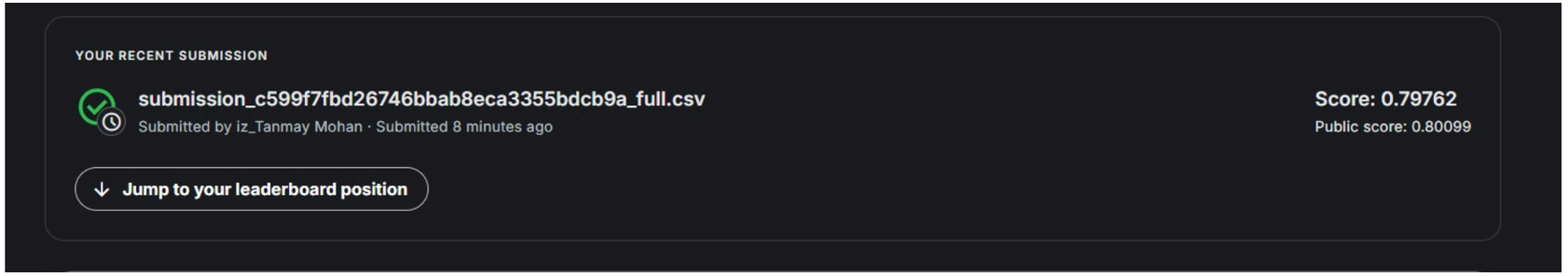
Our Novel features work!



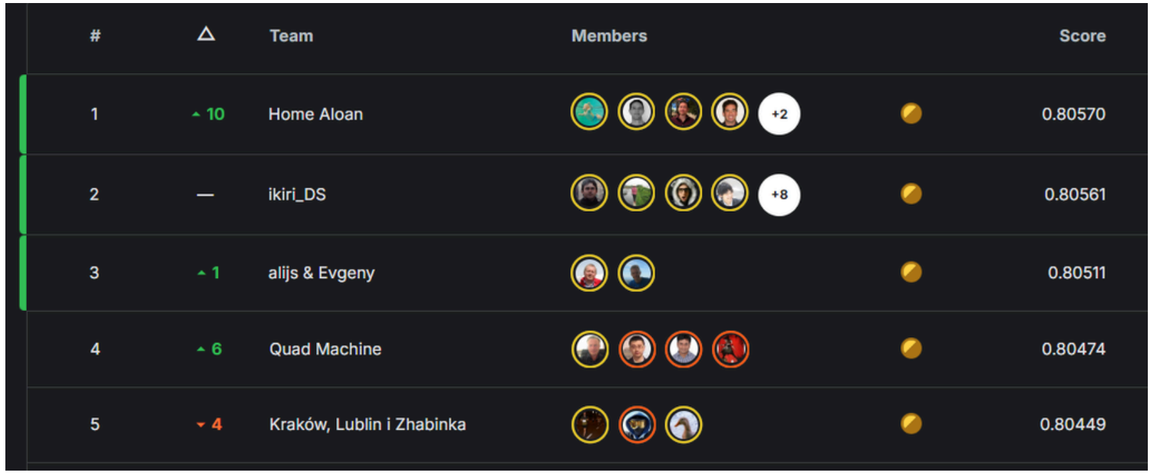



XGBoost (feature_importances_, gain; mean CV)
Novel in top-25: 15 / 25
KNN_TARGET_MEAN_K300
KNN_TARGET_MEAN_K500
KNN_TARGET_MEAN_K100
META_previous_application_MEAN
META_bureau_MEAN
EXT_SRC_MEAN
NAME_EDUCATION_TYPE_TE
BUREAU_BB_LAST_STATUS_SEV_MAX
CODE_GENDER_TE
META_bureau_MAX
EXT1_AGE_RATIO
CC_UTIL_P95_12M
ORGANIZATION_TYPE_TE
FLAG_OWN_CAR
META_installments_payments_MAX
BUREAU_DEBT_CREDIT_RATIO_MAX
OCCUPATION_TYPE_TE
META_previous_application_MIN
POS_DPD_MAX_6M
INST_1YR_LATE_FRAC
DEF_30_CNT_SOCIAL_CIRCLE
EXT_SOURCE_1_RESIDUAL
EXT_SRC_MIN
INST_2YR_LATE_FRAC
INST_LATE_FRAC
Novel (catalogue)
Standard
0.00
0.02
0.04
0.06
0.08
Importance (native) or fused score


Merged: mean of per-model percentile ranks (novel fusion / scale-free)
Novel in top-25: 15 / 25
KNN_TARGET_MEAN_K500
META_bureau_MEAN
META_previous_application_MEAN
KNN_TARGET_MEAN_K300
META_bureau_MAX
EXT_SRC_MEAN
EXT1_AGE_RATIO
BUREAU_DEBT_CREDIT_RATIO_MAX
META_previous_application_MIN
AMT_ANNUITY
KNN_TARGET_MEAN_K100
CODE_GENDER_TE
EXT_SOURCE_1_RESIDUAL
META_installments_payments_MAX
ORGANIZATION_TYPE_TE
DAYS_BIRTH
DOWNPAYMENT_INCOME_STRAIN
NAME_EDUCATION_TYPE_TE
EXT_SRC_MIN
OCCUPATION_TYPE_TE
ACTIVE_CREDIT_SUM_DECAYED
CC_UTIL_P95_12M
POS_CNT_INSTALMENT_FUTURE_MEAN
META_POS_CASH_balance_MAX
EXT2_EMPLOYED_PROD
Novel (catalogue)
Standard
0.0
0.2
0.4
0.6
0.8
1.0
Importance (native) or fused score



Position 170/7000+ teams



#	△	Team	Members	Score
1	▲ 10	Home Aloan	+2	0.80570
2	—	ikiri_DS	+8	0.80561
3	▲ 1	alijs & Evgeny		0.80511
4	▲ 6	Quad Machine		0.80474
5	▼ 4	Kraków, Lublin i Zhabinka		0.80449

Top 5 models from the same competition

Table VII. TEST RESULTS OF EACH CLASSIFIER. THE TIME IS TAKEN AS THE AVERAGE OF FIVE CONSECUTIVE TRAINING.

	Accuracy	*Precison*	*Recall*	*F1 score*	*AUC*	*Time(s)*
Lightgbm	0.734	0.181	0.657	0.284	0.772	213
XGBoost	0.849	0.235	0.39	0.293	0.735	2421
Logistic	0.645	0.544	0.641	0.589	0.692	251
SVM	0.658	0.543	0.637	0.586	0.688	[a]

a.Since the data set used in SVM training and testing is different from other classifiers, the time of SVM is not calculated here.

Table 1: Main performance comparison between tabular baselines, heterogeneous graph neural networks, and a hybrid ensemble on the Home Credit Default Risk dataset. Improvement is computed as relative ROC-AUC gain over logistic regression.

Category	Model	ROC-AUC ↑	PR-AUC ↑	ROC-AUC Improvement vs. Logistic
Tabular Baseline	Logistic Regression	0.7390	0.2160	–
Tabular Baseline	LightGBM (Strong Tabular)	0.7690	0.2540	+4.06%
Graph Neural Network	Contrastive Pretraining + Fine-tuning (GraphCL-style)	0.6804	0.1618	−7.93%
Graph Neural Network	Heterogeneous GraphSAGE (6-node hetero graph)	0.7400	0.2217	+0.14%
Graph Neural Network	**Proposed: Relation-Aware Attentive Heterogeneous GNN**	**0.7506**	**0.2291**	**+1.57%**

Why these metrics show that solution works?

1. The model separates risk levels
Defaulters are generally ranked above non-defaulters.
Stratified K-Fold for OOF predictions means class imbalance was taken into consideration while testing.

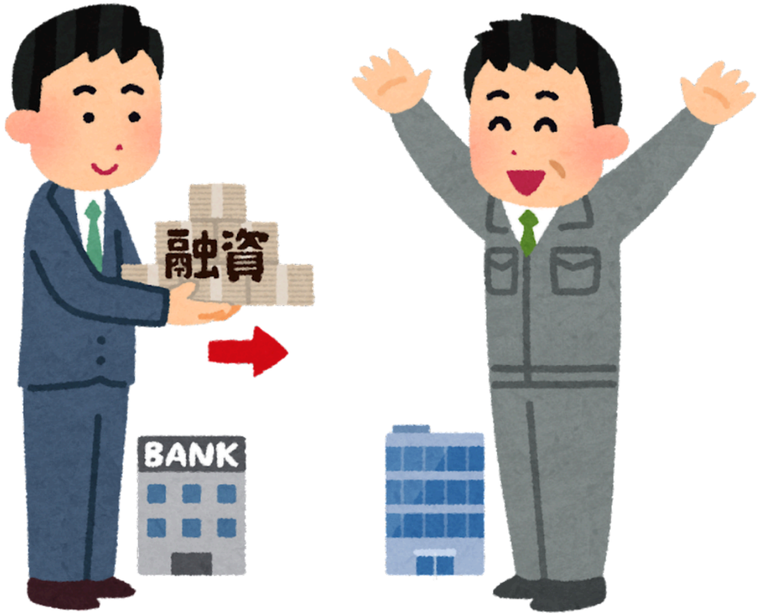
2. The model generalizes i.e. The Kaggle score is on unseen test data.

3. These metrics show that our solution works because ROC-AUC does not reward guessing the majority class. In this dataset, most applicants do not default, so accuracy alone could look high even for a weak model. ROC-AUC instead checks whether the model can consist
ently rank risk: a borrower who defaults should receive a higher risk score than a borrower who does not.

DEPLOYABILITY OF THE ML SOLUTION

For Private Banks & Financial Institutions - High Value

This model is directly deployable by:
Private Banks (HDFC, ICICI, Axis) - automated retail loan approval
NBFCs & Fintech platforms - real-time personal loan & BNPL scoring
Microfinance institutions - credit access for first-time borrowers



At Plaksha University - Limited

- Our model requires financial customer data that Plaksha does not generate
- Not applicable for direct campus deployment

Thank You