DATAVISION

Performance Metrics Report

November 16, 2025

Final Predictions (Submission)

Ensemble (Weighted Average)
Combine all 17 models based on validation AUC

Stacking (2 Meta-Models)
Train on meta features OOF predictionn

Model training (15 base mmodels)
LightGBM, XGBoost, CatBoost, MLP, TabM
x 3 feature sets

Feature Engineering (V1, V2, V3)
Create 6.000-7.000 features per version

Data cleaning $ Merging
standardize IDs, parse currency, handle missing values

Data loading 6 sources
Load CSV, Parquet, JSONL, XML, XLSX

Data Cleaning & Merging

Cleaning
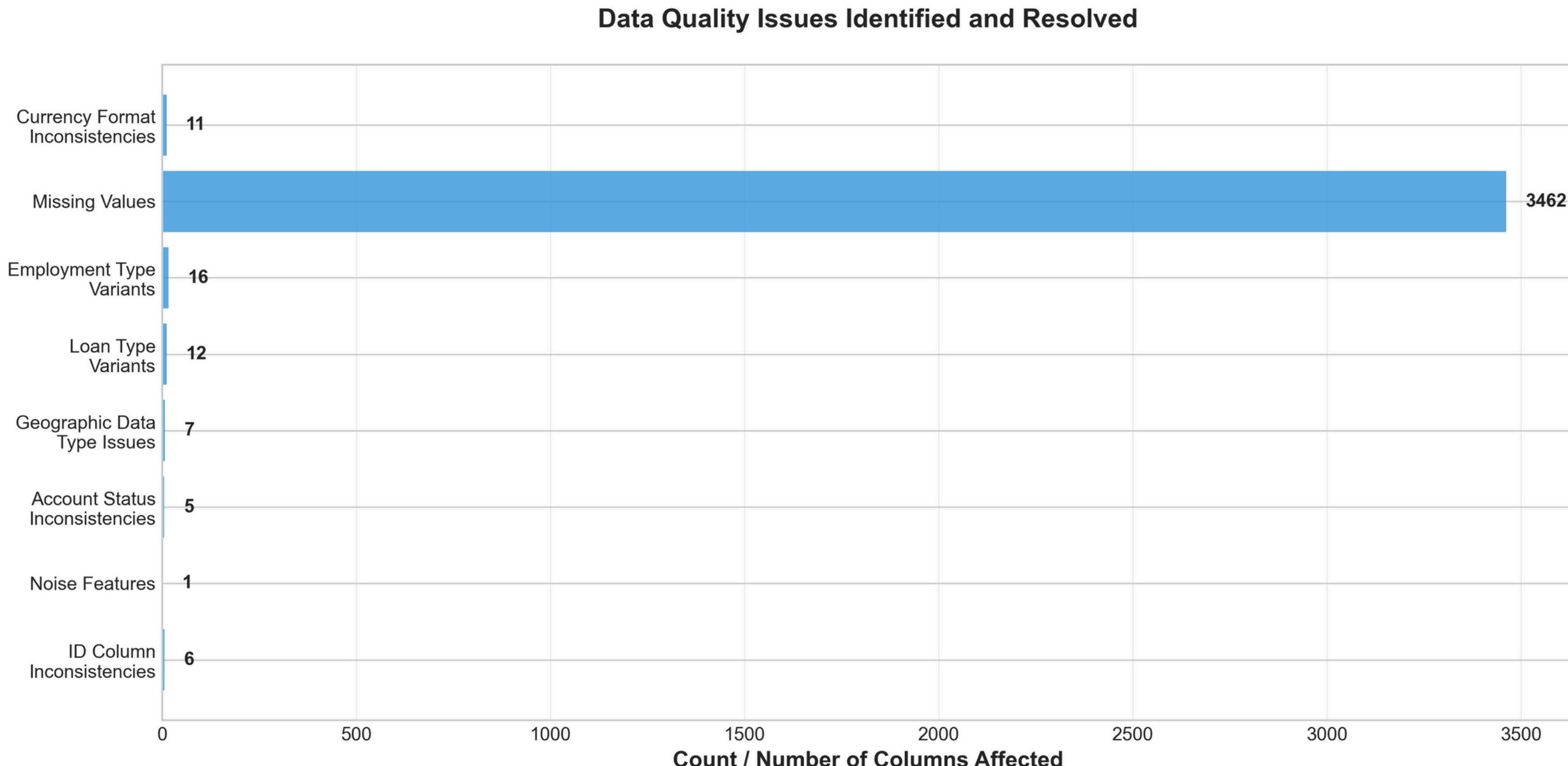
  parsing issues

    data types

      formats

Missing value strategy
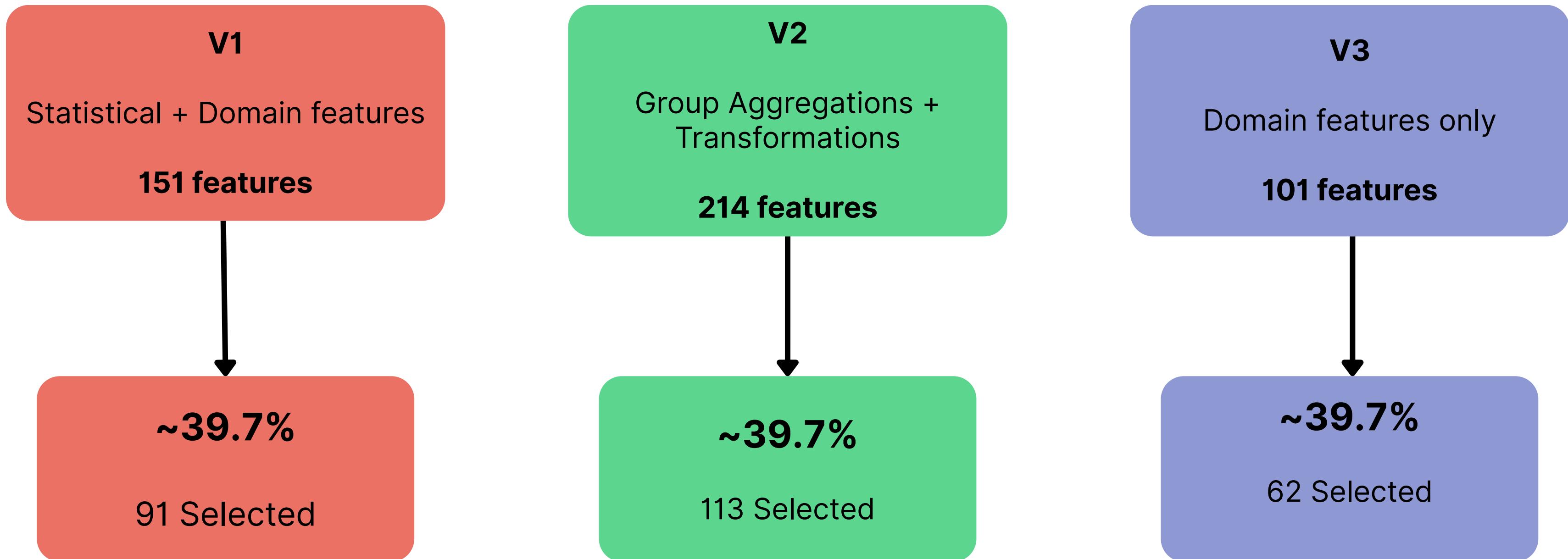
Median for continuous features

  Mode for categorical Features

Merging
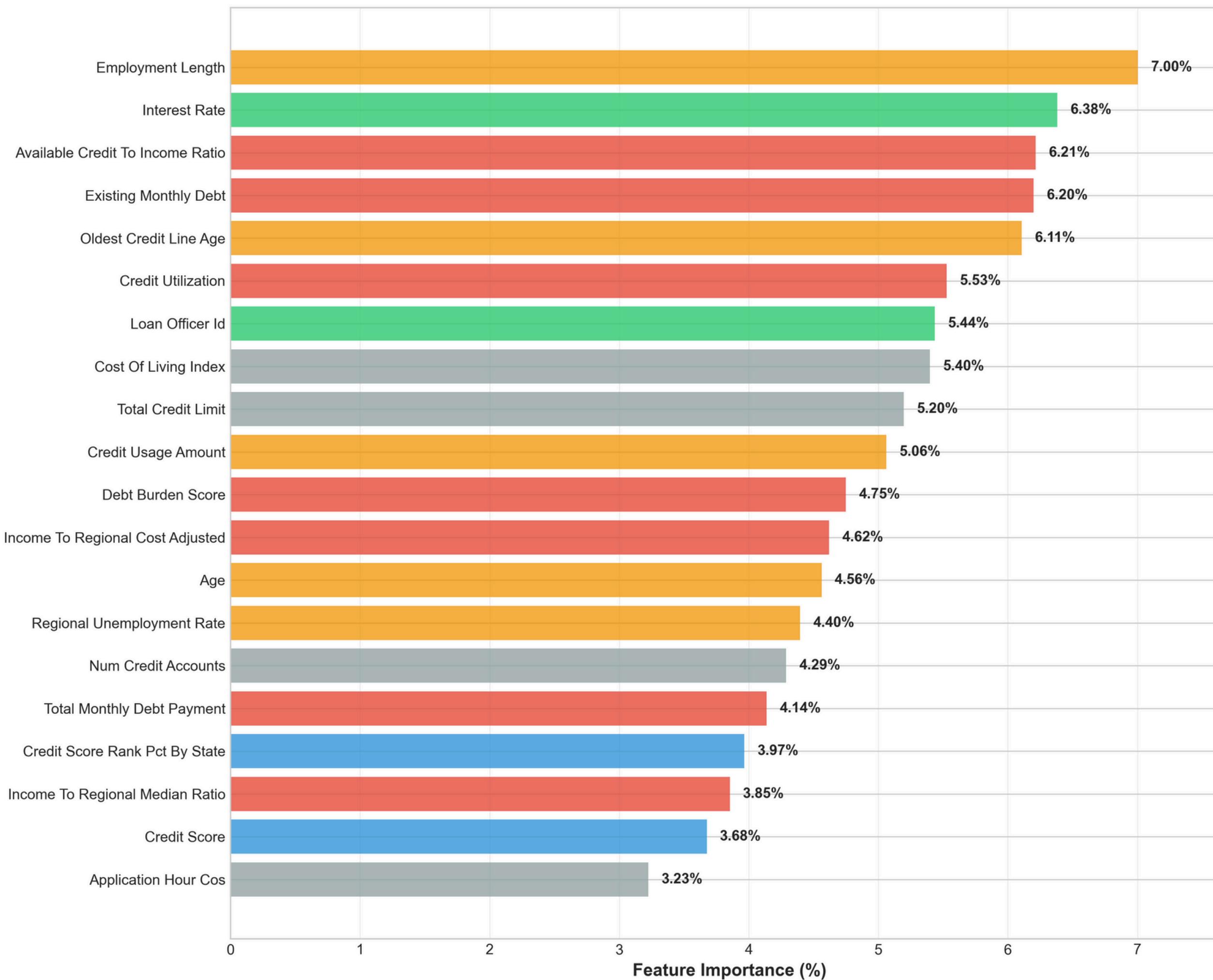
Combined datasets using *customer_id*

  Validated integrity after merging

Data Quality Issues Identified and Resolved

Issue	Count / Number of Columns Affected
Currency Format Inconsistencies	11
Missing Values	3462
Employment Type Variants	16
Loan Type Variants	12
Geographic Data Type Issues	7
Account Status Inconsistencies	5
Noise Features	1
ID Column Inconsistencies	6

FE & Selection
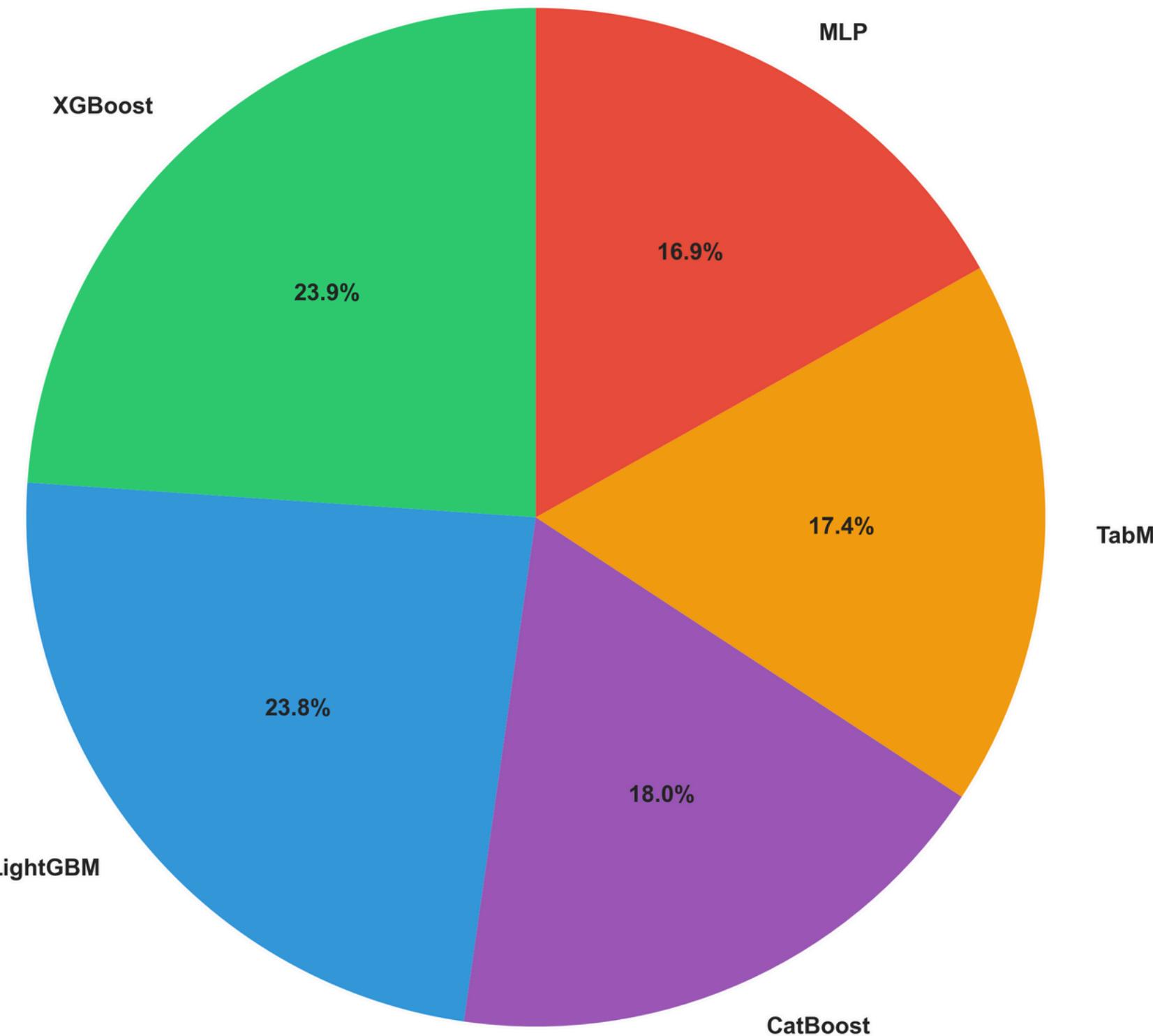
Top 20 Most Important Features

| Feature	Importance
Employment Length	7.00%
Interest Rate	6.38%
Available Credit To Income Ratio	6.21%
Existing Monthly Debt	6.20%
Oldest Credit Line Age	6.11%
Credit Utilization	5.53%
Loan Officer Id	5.44%
Cost Of Living Index	5.40%
Total Credit Limit	5.20%
Credit Usage Amount	5.06%
Debt Burden Score	4.75%
Income To Regional Cost Adjusted	4.62%
Age	4.56%
Regional Unemployment Rate	4.40%
Num Credit Accounts	4.29%
Total Monthly Debt Payment	4.14%
Credit Score Rank Pct By State	3.97%
Income To Regional Median Ratio	3.85%
Credit Score	3.68%
Application Hour Cos	3.23%

Feature Importance (%)

Model Diversity for Robustness

Ensemble Weights by Models

1. LightGBM
Fast gradient boosting tree model.

2. XGBoost
High-accuracy boosting algorithm.

3. CatBoost
Boosting optimized for categorical data.
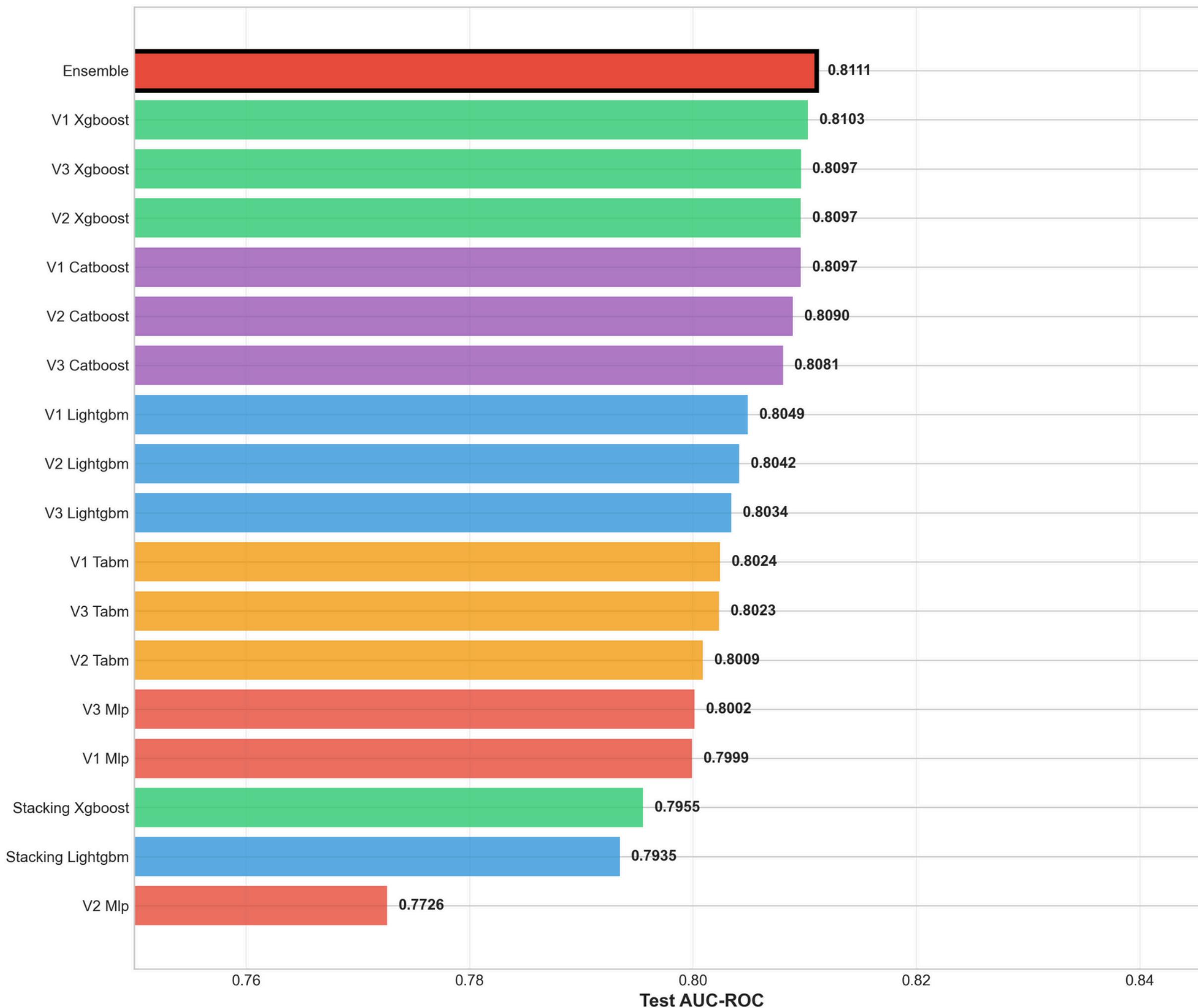
4. MLP
Artificial neural network model.
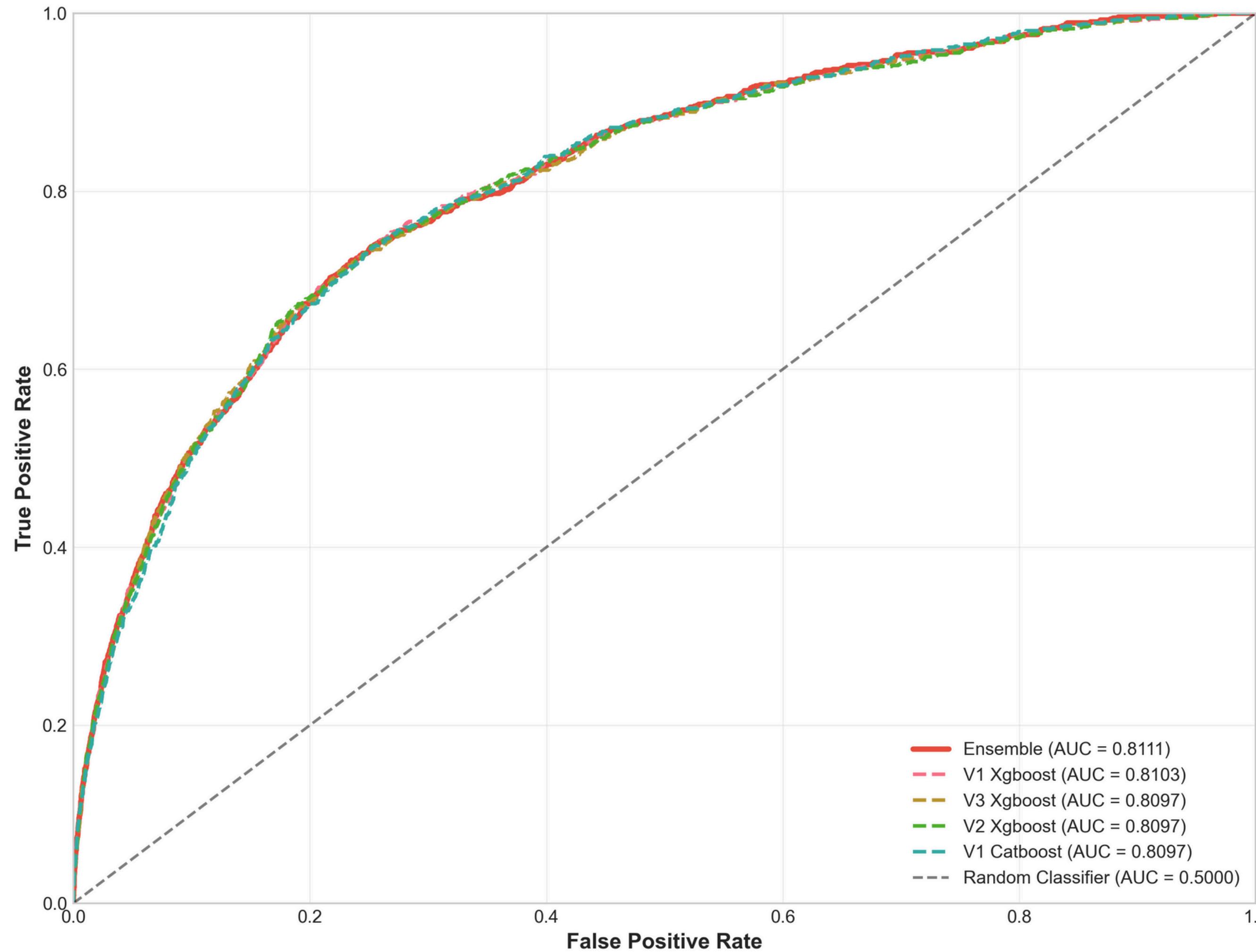
5. TabM
Model suited for tabular data.
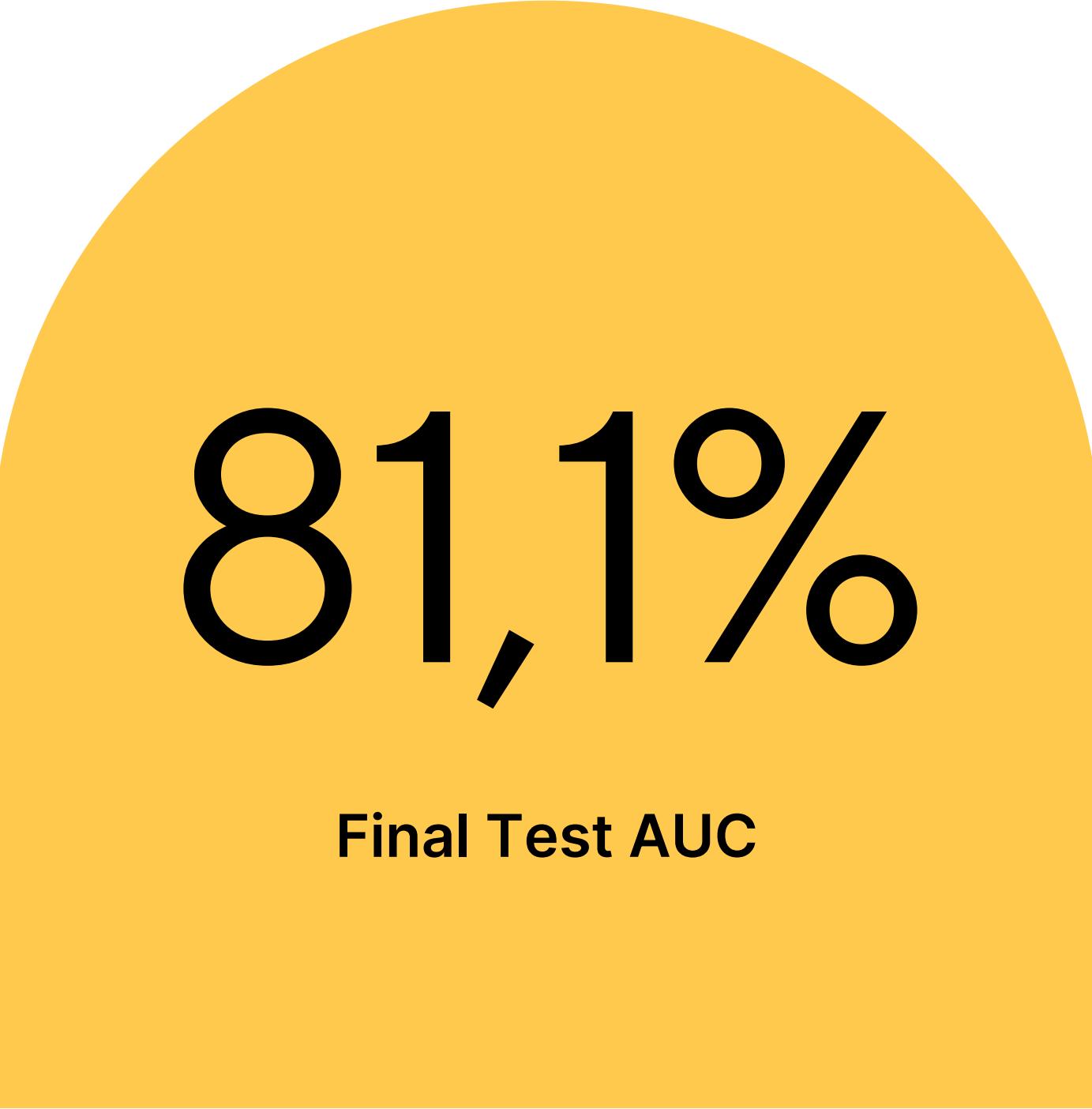
Each model trained on 3 different feature sets (V1,V2,V3)

(15 base models + 2 meta models)

Model Performance Comparison (All 17 Models + Ensemble)

Model	Test AUC-ROC
Ensemble	0.8111
V1 Xgboost	0.8103
V3 Xgboost	0.8097
V2 Xgboost	0.8097
V1 Catboost	0.8097
V2 Catboost	0.8090
V3 Catboost	0.8081
V1 Lightgbm	0.8049
V2 Lightgbm	0.8042
V3 Lightgbm	0.8034
V1 Tabm	0.8024
V3 Tabm	0.8023
V2 Tabm	0.8009
V3 Mlp	0.8002
V1 Mlp	0.7999
Stacking Xgboost	0.7955
Stacking Lightgbm	0.7935
V2 Mlp	0.7726

Test AUC-ROC

ROC Curve Comparison - Ensemble vs Top Individual Models

Ensemble (AUC = 0.8111)
V1 Xgboost (AUC = 0.8103)
V3 Xgboost (AUC = 0.8097)
V2 Xgboost (AUC = 0.8097)
V1 Catboost (AUC = 0.8097)
Random Classifier (AUC = 0.5000)
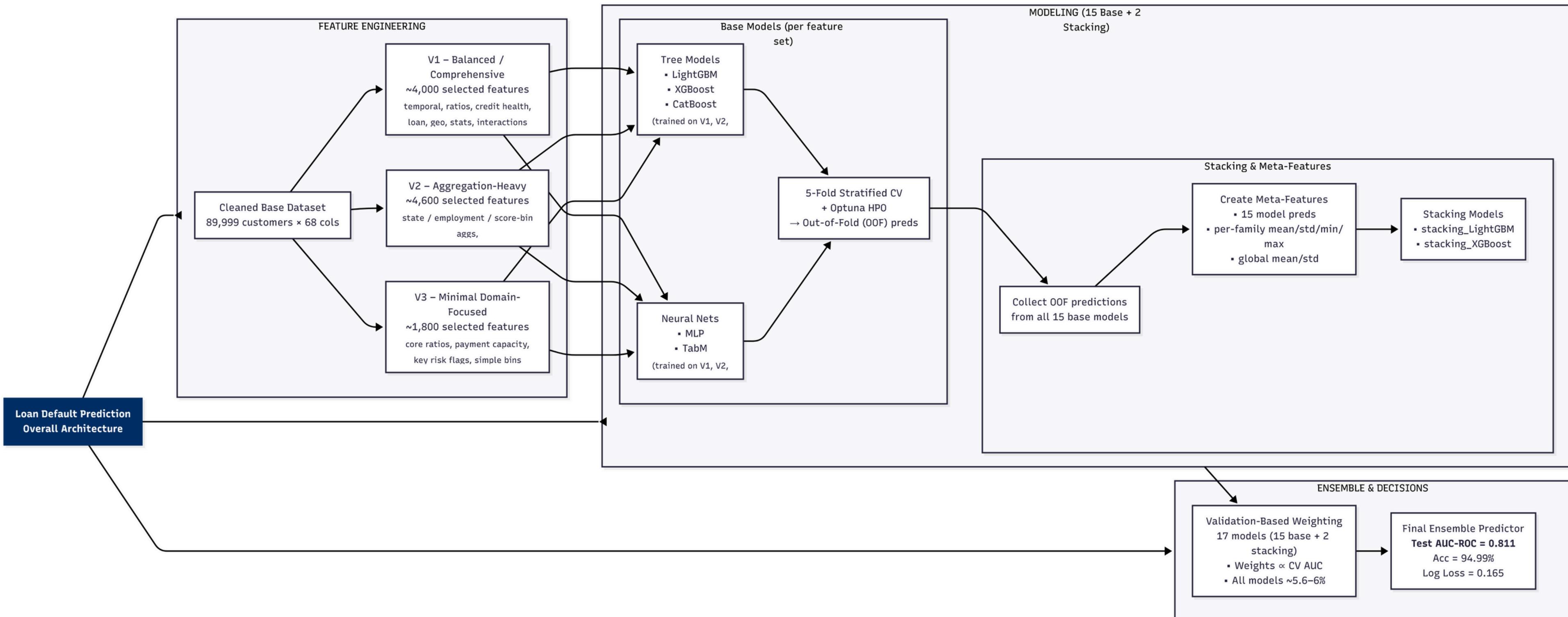
81,1%

Final Test AUC

✓ **Final Test Set Performance:** AUC-ROC = 0.811064

✓ **Cross-Validation Performance:** AUC-ROC = 0.805022

✓ **Test Set Accuracy:** 94.99%

✓ **Test Set Log Loss:** 0.165026

Loan Default Prediction Overall Architecture

FEATURE ENGINEERING

Cleaned Base Dataset
89,999 customers × 68 cols

V1 – Balanced / Comprehensive
~4,000 selected features
temporal, ratios, credit health, loan, geo, stats, interactions

V2 – Aggregation-Heavy
~4,600 selected features
state / employment / score-bin aggs,

V3 – Minimal Domain-Focused
~1,800 selected features
core ratios, payment capacity, key risk flags, simple bins

MODELING (15 Base + 2 Stacking)

Base Models (per feature set)

Tree Models
▪ LightGBM
▪ XGBoost
▪ CatBoost
(trained on V1, V2,

Neural Nets
▪ MLP
▪ TabM
(trained on V1, V2,

5-Fold Stratified CV
+ Optuna HPO
→ Out-of-Fold (OOF) preds

Stacking & Meta-Features

Collect OOF predictions from all 15 base models

Create Meta-Features
▪ 15 model preds
▪ per-family mean/std/min/max
▪ global mean/std

Stacking Models
▪ stacking_LightGBM
▪ stacking_XGBoost

ENSEMBLE & DECISIONS

Validation-Based Weighting
17 models (15 base + 2 stacking)
▪ Weights ∝ CV AUC
▪ All models ~5.6–6%
Acc 94.99%
Log Loss = 0.165

Final Ensemble Predictor
Test AUC-ROC = 0.811
Acc = 94.99%
Log Loss = 0.165

Our Team

Javohir - Data Analyst & Engineer

Dilyorbek - AI/ML developer

Maruf - Data Engineer

Mirjahon - Data Engineer & Analyst

Aziz - AI/ML Engineer

Thank you