Home Credit Default Risk Prediction

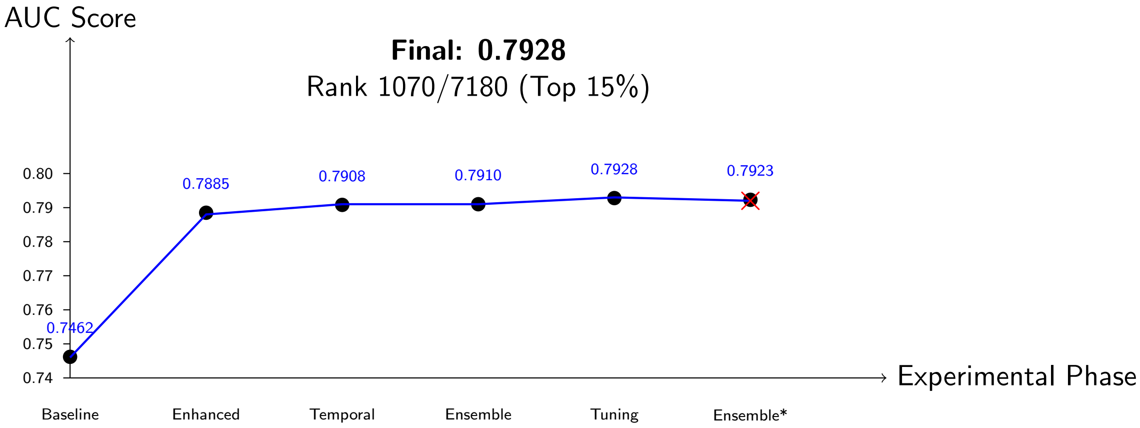
A Journey Through Feature Engineering and Model Optimization

YANG, Kuo

SID: 21020376

November 19, 2025

Project Journey Overview



Key Question: What worked, what didn't, and why?

Starting Point: The Baseline

Configuration:

- Model: LightGBM (default hyperparameters)
- Features: Basic aggregations from auxiliary tables
  - Mean, max, min, sum
  - Simple derived features (age, income ratios)
- Validation: 5-fold stratified CV
- Total features: 129

Result: Private Score = 0.7459

Rank: 5847/7180

Solid foundation with proper validation strategy

Phase 1: Enhanced Aggregation Features

Added 155 new features (129 $\rightarrow$ 284):

1. External Source Interactions

$$\text{MEAN} = \frac{1}{3}\sum_{i=1}^{3} \text{EXT}_i$$

$$\text{WEIGHTED} = 0.5 \cdot E_1 + 0.3 \cdot E_2 + 0.2 \cdot E_3$$

2. Debt Analysis

$$\text{RATIO} = \frac{\sum \text{DEBT}}{\sum \text{CREDIT}}$$

Feature engineering dominated all other improvements

3. Payment Behavior

$$\text{LATE\%} = \frac{\#\ \text{late payments}}{\#\ \text{total payments}}$$

Result: +0.042 AUC

57% of total gains!

Phase 2: Temporal Features

Static vs. Dynamic Behavior

- **Problem:** Aggregations miss behavioral changes
- **Solution:** Compare recent vs. historical patterns

Added 76 temporal features (284 $\rightarrow$ 360):

- **Bureau Balance Trends:** $\text{Recent}_{6m} - \text{Old}_{>12m}$
- **Spending Velocity:** $\frac{\text{Recent spending} - \text{Old spending}}{\text{Old spending}}$
- **Payment Delay Evolution:** 2nd half delays − 1st half delays

Result: +0.002 AUC

Modest but consistent—captures behavioral dynamics

Phase 3: The Ensemble Experiment

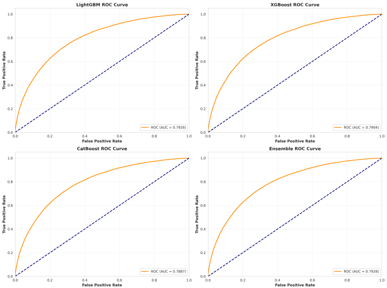
Conventional Wisdom: More models = Better predictions

Three Models:

- LightGBM
- XGBoost
- CatBoost

Optimization:

- Grid search over **232 weight combinations**
- Optimal: $P = 0.75P_{LGB} + 0.15P_{XGB} + 0.10P_{Cat}$



All models show similar AUC (0.7887-0.7928). Ensemble provides marginal improvement.

Pre-Tuning Result: +0.0002 AUC

The Tuning Paradox

Hyperparameter Optimization: Optuna with Bayesian search (100 trials)

Configuration	Private Score	vs. Pre-Tuning	Rank
Single LightGBM (tuned)	**0.7928**	+0.0018	**1070**
Ensemble (tuned)	0.7923	-0.0013	1327

Ensemble got WORSE after tuning!

Model Correlation Matrix (Post-Tuning):

	LightGBM	XGBoost	CatBoost
LightGBM	1.00	0.976	0.968
XGBoost	—	1.00	0.981

Models became too similar—lost diversity

Why Ensembles Failed Post-Tuning

Two Key Mechanisms:

1. Reduced Model Diversity

- Hyperparameter optimization pushed all models toward similar optima
- LightGBM & XGBoost both converged: `max_depth` ≈ 9, `lr` ≈ 0.028
- High correlation ($\rho > 0.97$) $\rightarrow$ same mistakes

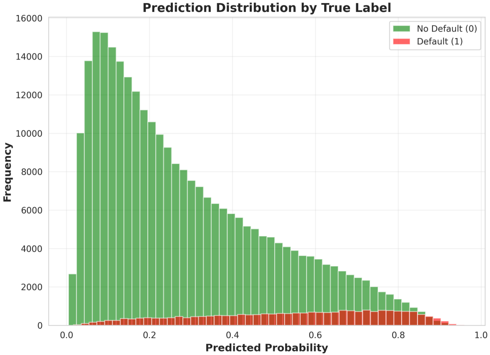
2. Bias-Variance Tradeoff Shift

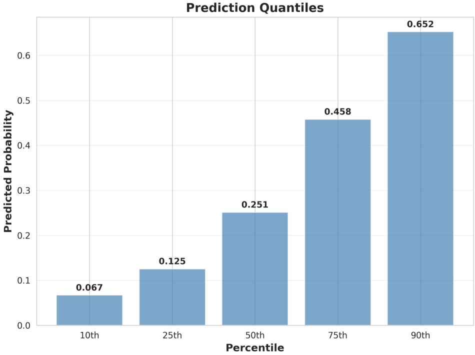
- Pre-tuning: High bias $\rightarrow$ averaging reduces variance
- Post-tuning: Near-optimal $\rightarrow$ averaging adds unnecessary smoothing

Competition Intensity:

- Tuned ensemble only 0.0005 worse than single model
- But cost nearly **300 ranks** (1070 $\rightarrow$ 1327)
- Every fraction of a point matters!

What the Model Actually Learned

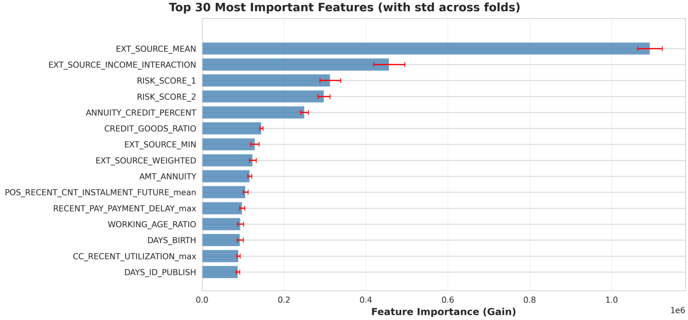




Key Observations:

- **Well-separated distributions:** Clear distinction between defaults (red) and non-defaults (green)
- **Conservative predictions:** Median = 0.25 (appropriate for financial inclusion)
- **High-risk tail:** Only 10% above 0.65 → focus manual review efficiently

Feature Importance Analysis



- **EXT_SOURCE_MEAN dominates:** Massive importance gap
- **Engineered features win:** Income interaction (#2) and risk scores (#3) validate domain design
- **Temporal signals matter:** Recent behavior features appear throughout top-15
- **Ethical concern:** Bureau scores aren't available for unbanked populations

Alternative behavioral signals can partially compensate

Performance Progression Summary

Phase	Features	Score	ΔAUC	Rank
Baseline	129	0.7462	—	5847
Enhanced Aggregation	284	0.7885	+0.042	3030
Temporal Features	360	0.7908	+0.002	2766
Pre-Tuning Ensemble	360	0.7910	+0.0002	2650
Hyperparameter Tuning	360	**0.7928**	+0.0018	**1070**
Post-Tuning Ensemble	360	0.7923	-0.0005	1327

Contribution Breakdown:

- Feature engineering: **57%** of total gains
- Hyperparameter tuning: **24%** of total gains
- Ensemble learning: ≈ 0 **%** after tuning

Cumulative: +0.047 AUC → Rank 5847 to 1070

Key Lessons Learned

1. Feature Quality > Model Complexity

Domain-informed feature engineering outweighed all algorithmic improvements

2. Ensemble Learning Has Diminishing Returns

When individual models are well-tuned, ensembles add complexity without gains

3. Validation Strategy is Critical

Consistent 5-fold stratified CV prevented chasing validation noise

Simpler often beats complex

Limitations & Future Work

Acknowledged Limitations:

- **Temporal validation:** Features may incorporate post-application information
  - Need rigorous temporal cutoffs for production deployment
- **Data equity:** Heavy reliance on credit bureau scores
  - Perpetuates exclusion of unbanked populations
  - Alternative signals help but more work needed

Future Directions:

- **Deeper feature engineering:** Learn from Kaggle discussion forum
  - Successful participants share many effective feature calculations
  - Rich source of domain insights
- **Complex models (neural networks, deep learning):**
  - Could further improve scores and rankings
  - But likely less efficient than discovering better features

Feature engineering remains the highest leverage activity

Conclusion

Core Findings:

- Thoughtful feature engineering beats algorithmic complexity
- Simple, well-tuned models often outperform complex ensembles
- Faster inference + better interpretability + same performance = win

Sometimes, simpler is better.

Thank you!