Predicting Corporate Bankruptcy Using Ensemble Machine Learning Models

Team 404coder (IIT ISM Dhanbad)

November 4, 2025

1 Introduction

Predicting corporate bankruptcy is a high–stakes imbalanced classification problem. Only ∼7% of companies in the dataset fail, making macro–F1 the appropriate evaluation metric. Our objective was to maximize macro–F1 on the unseen leaderboard data.

Initial experiments with individual models plateaued near 0.50 F1 on leaderboard. We identified feature leakage and model instability due to naive validation strategy and corrected it.

2 Dataset

The training set consisted of company financial ratios and industry identifiers.

Features

- Numerical financial predictors: X1–X18 (profitability, leverage, liquidity, efficiency ratios)
- Categorical: Industry Division (post–processed); MajorGroup removed due to leakage
- Target: `status_label` (Alive / Failed)

Data Cleaning

- Removed: `Unnamed:0`, `company_name`, `fyear`, `MajorGroup` (leak)
- Rare levels of `Division` merged to "Other"
- Ordinal encoding for tree models; CatBoost ingested raw categorical
- Train/validation split: stratified 80/20

3 Models

Three gradient–boosting models trained on cleaned data:

- CatBoost Classifier
- LightGBM Classifier

- XGBoost Booster (DMatrix, XGBoost v3.x API)

Early stopping was used for all models. CatBoost hyperparameters were inherited from prior Optuna tuning; LightGBM and XGBoost tuned pragmatically for stability and minority recall.

4 Ensemble Strategy

To avoid overfitting from stacking and excessive search time, we used a weighted probability ensemble.

Weights and threshold optimized using Optuna to maximize macro–F1:

$$w = \arg\max F_1 \left(y, \mathbf{1} \left(\sum_i w_i p_i > t \right) \right)$$

Final Ensemble Weights

$$w_{\text{CatBoost}} = 0.175, \quad w_{\text{LightGBM}} = 0.614, \quad w_{\text{XGBoost}} = 0.211$$

Optimal prediction threshold:

$$t = 0.2097$$

5 Results

Validation Performance

Macro–F1 on validation:

$$\text{F1}_{macro} = \mathbf{0.6562}$$

Model	CatBoost	LightGBM	XGBoost
Individual Macro–F1	0.4745	0.5407	0.5437
Ensemble Macro–F1	**0.6563**	–	–

Table 1: Validation scores before and after ensembling

Interpretation

- CatBoost performance dropped after removing leakage — expected and desired
- LightGBM contributed the dominant signal
- Ensemble significantly increased minority recall without destroying precision

6 Conclusion

A carefully constructed weighted ensemble of boosting models, combined with leakage removal and threshold tuning, substantially improved macro–F1 from ∼0.50 (leaderboard baseline) to a validated score of **0.656**.

Future work includes:

- Temporal validation to mitigate dataset drift
- Financial ratio engineering and winsorization
- Pseudo–labeling if leaderboard gap persists

Code Repository

The complete training and inference pipeline is available at:
`https://github.com/Sukrat-Singh/p2p-hackathon`