AIDI-2004-02 AI IN ENTERPRISE SYSTEMS

LAB04

SUBMITTED BY
AROMAL GIGI
(100990951

Company Bankruptcy Prediction- Key Decision

1) Choosing Initial Models

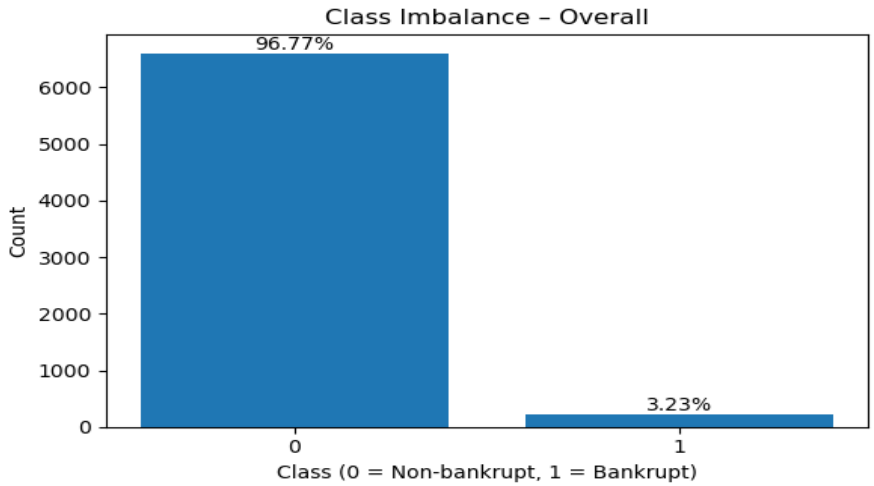
- Benchmark: Logistic Regression → interpretable, quick to run
- Complex models: Random Forest & XGBoost → handle non-linearities, strong tabular performance
- Skip clustering: Labels available; supervised outperforms unsupervised here
- Benchmark enables fair comparison of complex vs. simple models

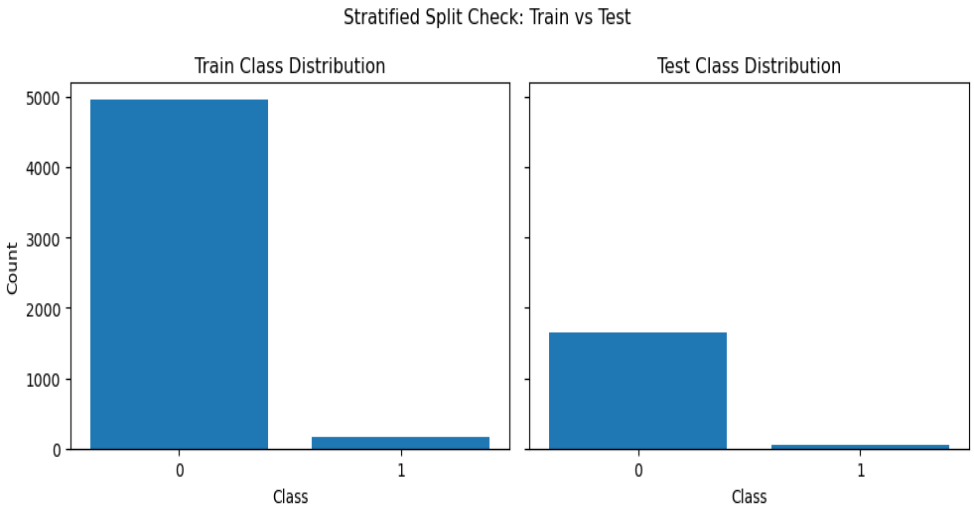
2) Data Pre-processing

- LR: Requires scaling for stability and convergence
- Trees: No scaling needed; unaffected by feature scale
- Encoding: Not required; all features numeric
- Pipeline: Separate per model to prevent leakage

3) Handling Class Imbalance

- Detected: Significant minority class imbalance (bankruptcy rate ~3.23%)
- SMOTE: Boosts minority samples; risk of synthetic noise if overused
- Class weights: Balances loss function without altering data
- Stratify: Maintains ratio in train/test and CV splits





4) Outlier Detection & Treatment

- Keep extremes: May signal high-risk behavior; valuable predictive info
- Remove only errors: NaN, inf, or impossible values
- Winsorize noise: Cap extreme non-informative outliers
- Preserve minority class patterns while reducing harmful noise

5) Train/Test Sampling Bias (PSI / Corr)

- Why test?: Different distributions → optimistic CV, degraded deployment
- Do: Compute PSI per feature; flag PSI > 0.25 as large shift
- Act: Re-split, re-weight, or drop highly shifted features
- Rationale: Detect instability before deployment

6) Data Normalization

- Manual for LR: StandardScaler for stable coefficients and convergence
- Not needed for trees: Splits are rank-based, scale-independent
- Consistency: Fit scaler on train only; apply to validation/test
- Rationale: Prevents leakage and avoids unnecessary transformations for tree models

7) Testing for Normality

- Trees: Not affected by non-normal features; no action needed
- LR: Strong skew may affect linearity; can log-transform skewed positive features

- Apply only if: CV shows performance improvement after transformation
- Rationale: Avoid unnecessary transformations unless they provide measurable gains

8) Dimensionality Reduction (PCA)

- Skip for trees: Already handle collinearity; no gain from PCA
- Optional for LR: Can reduce multicollinearity and noise if VIF high
- Pros: Reduces dimensionality, may lower overfitting risk, decorrelates features
- Cons: Loses original feature meaning, may obscure important domain patterns

9) Feature Engineering Choices

- Minimal additions: Only theory-driven interactions or transformations
- Dataset already rich: Contains many predictive financial ratios
- Avoid leakage: No features derived from target or future info
- Rationale: Maintain interpretability and prevent overfitting with unnecessary features

10) Testing and Addressing Multicollinearity

- Detection: Use correlation matrix and VIF (VIF > 10 = concern)
- Impact: Inflates LR coefficient variance, reduces interpretability
- Action: Drop one of highly correlated pairs for LR; no change for trees
- Rationale: Stabilizes LR while keeping predictive power intact

12) Feature Selection Methods

- Method: Correlation filter ($|r| > 0.95$) + model-based importance from XGBoost
- Too many: Risk of overfitting and noise inclusion
- Too few: Miss important signals, underfit the model
- Rationale: Retain only impactful, non-redundant features to balance performance and generalization

13) Hyperparameter Tuning Methods

- Approach: Random Search for broad coverage, then Optuna for fine-tuning
- Reason: Random Search quickly explores large parameter space efficiently

- Optuna: Refines top configurations with fewer iterations
- Rationale: Balances search efficiency with performance optimization for complex models

14) Cross-Validation Strategy

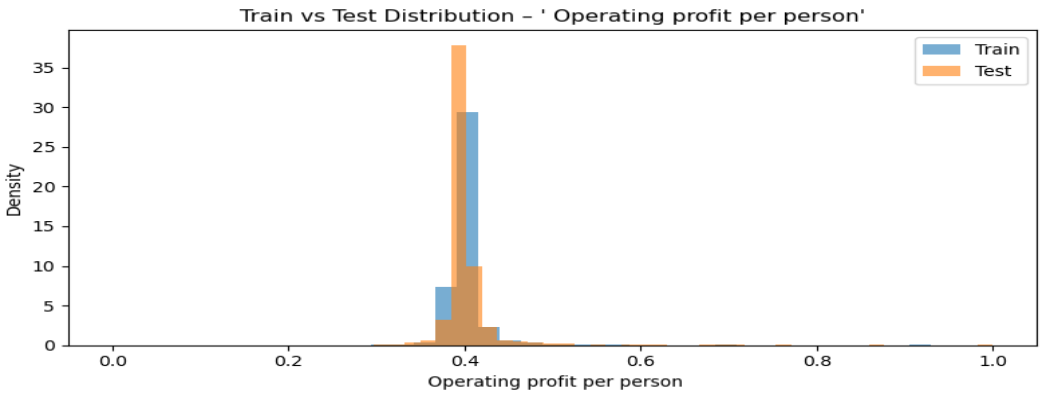
- Selected: Stratified K-fold
- Reason: Preserves class imbalance ratio across folds
- Benefit: More reliable performance estimates for minority class
- Rationale: Reduces metric variance and improves generalization assessment

15) Evaluation Metrics Selection

- Selected: PR-AUC (primary), Recall@target-precision, F1/F2, ROC-AUC
- Reason: PR-AUC focuses on positive class quality in imbalanced data
- Approach: Use predict_proba(X) for probability-based metrics
- Rationale: Balances ranking quality (ROC-AUC) with actionable recall/precision trade-offs

16) Evaluating Drift and Model Degradation

- **Selected:** PSI between train and test sets
- **Reason:** Detects distribution changes that can reduce model accuracy
- **Benefit:** Confirms all features < 0.1 PSI → stable; no features > 0.25 → no large drift
- **Rationale:** Ensures model is trained on data representative of future inputs



17) Interpreting Model Results and Explainability

- Selected: SHAP for tree models, coefficients for Logistic Regression
- Reason: SHAP provides both global and local interpretability
- Benefit: Transparent reasoning for each prediction, supports audits
- Rationale: Meets regulatory requirements and builds stakeholder trust