Bankruptcy Prediction

Develop a binary classification model to predict if a firm will file for bankruptcy.

PURDUE UNIVERSITY®

Introduction

The Challenge: Accurately classify firms as "Solvent" or "Bankrupt" using noisy financial data.

Key Obstacles:

High dimensionality (many features).	Missing data.	Complex, non-linear relationships between financial ratios.

Our Strategy: Maximize model diversity by stacking **Gradient Boosting** with **Evolutionary Algorithms**.

PURDUE UNIVERSITY

Imputation Strategy

Why two strategies? Different models process math differently.

Strategy A: XGBoost & LightGBM (-999)

Missing values replaced with -999.

Why: Tree models treat this as a distinct category/branch.

Strategy B: Evolutionary XGB (Median)

Missing values replaced with Median.

Why: EXGB performs math (e.g., Feature A + Feature B). Using -999 would create massive outliers (e.g., -994) and destroy linear patterns.

PURDUE UNIVERSITY

Feature Engineering & Selection

The Problem: Raw financial data often hides the signal.

The Solution: "Brute Force" Interaction Generation.

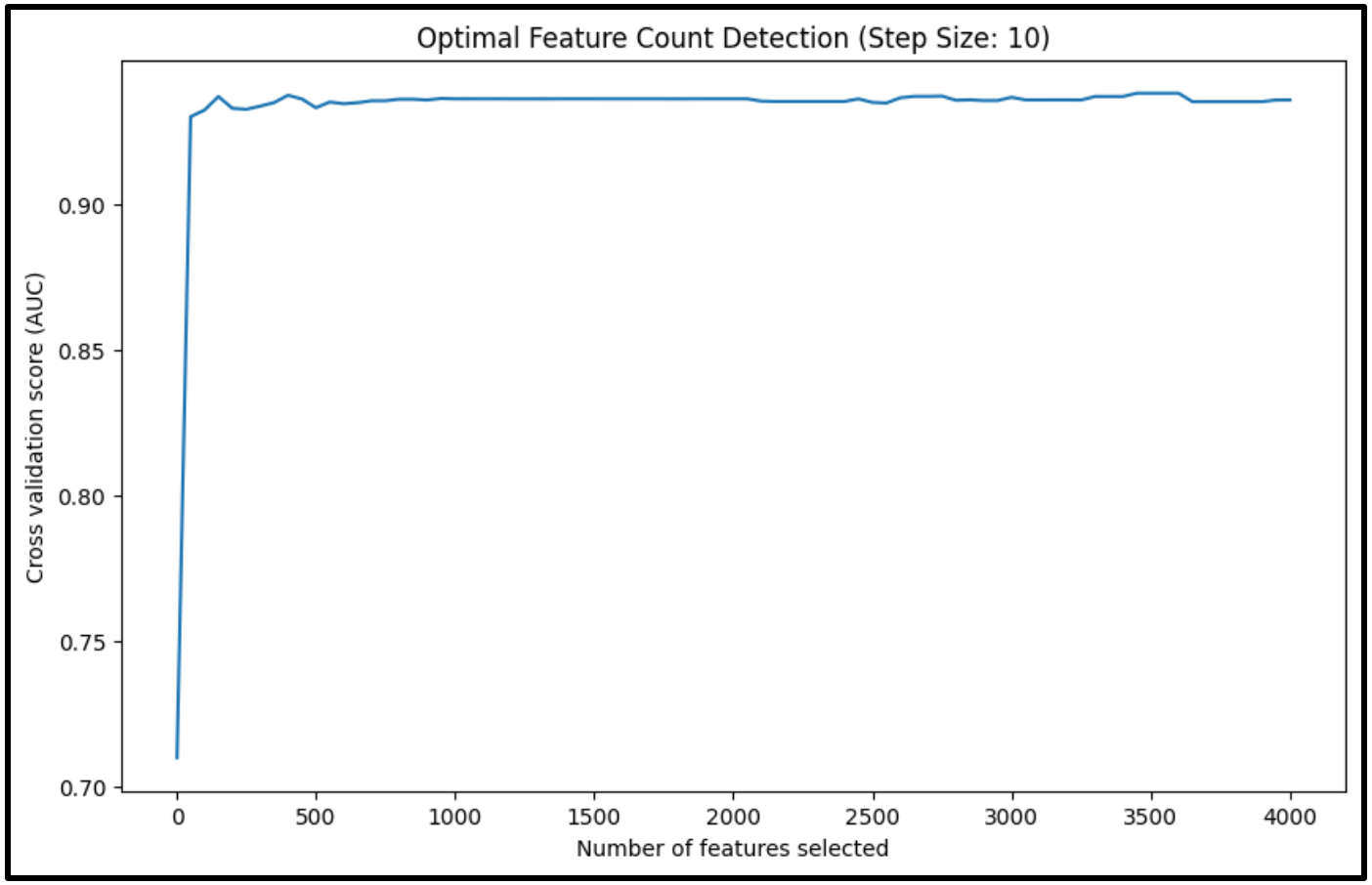
- Generated **16,000+** features using arithmetic (Add, Sub, Mult, Div).
- Real examples that became top features include **(Current Assets – Current Liabilities) × Inventory Turnover** (liquidity stress indicator), as well as **Total Debt ÷ (Net Income + Depreciation)** (a variant of interest coverage).

The Filter: "Scout Model" Approach.

- Trained a fast XGBoost model to calculate Information Gain.
- Reduced 16,000 features down to the top features contributing the most.
- That's a 98% reduction in dimensionality while keeping almost all the predictive power.

PURDUE UNIVERSITY

Feature Engineering & Selection



- This curve shows the cross-validation AUC as we add more scout-ranked features.

- Performance increases sharply at first, then plateaus around **300–350 features**, after which additional features introduce noise and overfitting.

- Therefore, **300–400 features became our optimal range**.

PURDUE UNIVERSITY

Model Architecture

The Structure: A 3-Stage Stacking Pipeline.

Level 1 (Base Learners - The "Bagging" Layer):

- **10 XGBoost Models:** Trained on 10 distinct random seeds.
- **10 LightGBM Models:** Trained on 10 distinct random seeds.
- **Validation:** Every single seed used **5-Fold Stratified Cross-Validation**.
- *(Total: 100 individual training runs just for this layer)*.

Level 2 (Evolutionary): Evolutionary XGBoost (EXGB).

Level 3 (Meta-Learner): Logistic Regression Stacker.

PURDUE UNIVERSITY

EVOLUTIONARY XGBOOST

What is it? A genetic algorithm wrapper around XGBoost.

Configuration: 10 Learners.

The Process:

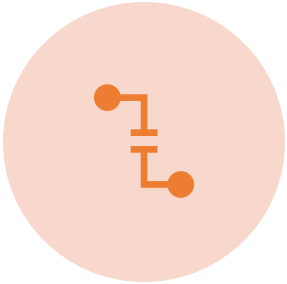
- Iterates through **100 rounds**.
- In every round, it generates **60 new synthetic features**.
- **Survival of the Fittest:** Only keeps features that improve the AUC score; discards the rest.

PURDUE UNIVERSITY

Challenges

Model Choice
Deciding the right model

Meta Learner
Deciding what method to choose (Hill Climbing, Logistic Regression, etc.)



Hyperparameter Tuning
Adjusting the hyperparameters to reduce overfitting

Feature Selection
Adjusting the number of features to reduce noise

Key Learnings

Baseline: Simple Logistic Regression (AUC: ~0.85).

Single XGBoost: Good, but high variance (AUC: ~0.91).

Random Forest: Struggled with the high dimensionality.

Neural Networks: Overfitted due to the dataset size.

Conclusion: The **Stacking Ensemble** provided the best balance of bias and variance.

PURDUE UNIVERSITY

Thank You

PURDUE UNIVERSITY®