

Three Years Before Bankruptcy: Multi-Model Early Warning of Corporate Financial Distress

Abstract

Corporate bankruptcy is a culmination of prolonged financial deterioration rather than a sudden event. This study investigates the feasibility of predicting corporate distress three years prior to failure, utilizing a comprehensive dataset of U.S. listed firms from 1999 to 2018 (N=78,682). We employ and compare three distinct modeling approaches—Logistic Regression, XGBoost Tree Model, and Deep Neural Networks (DNN)—to identify early warning signals within 18 key financial indicators. Addressing the challenge of severe class imbalance (6% failure rate), our results demonstrate that financial distress follows a progressive trajectory, with *Retained Earnings* and *Market Value* showing significant divergence well before default. Among the evaluated models, XGBoost emerges as the superior classifier, effectively capturing non-linear risk patterns and achieving the optimal trade-off between Recall and Precision. Conversely, the Deep Neural Network underperforms in this tabular setting, highlighting the limitations of complex architectures for structured financial data. This paper confirms that a three-year early warning system is not only statistically viable but also offers practical utility for stakeholders to mitigate default risks proactively.

1. Introduction

Corporate bankruptcy is rarely an instantaneous event; rather, it is the culmination of prolonged financial deterioration. Significant distress signals—such as declining profitability and structural imbalances—often manifest years before a firm's actual default. For stakeholders ranging from commercial banks to equity investors and regulators, the ability to identify these "pre-bankruptcy" trajectories 1 to 3 years in advance is critical for risk mitigation and capital preservation. This project addresses a central question: **How do financial patterns evolve in the three years preceding failure, and can these early signals be effectively captured by predictive models?**

To investigate this, we analyze a comprehensive dataset of U.S.-listed companies spanning from 1999 to 2018. The dataset comprises approximately 78,000 firm-year observations across nearly 9,000 unique firms. By tracking 18 key financial indicators derived from corporate balance sheets and income statements, we reconstruct the financial path of firms as they approach bankruptcy. A defining characteristic of this dataset is its severe class imbalance, with failed firms constituting only about 6% of the sample—a realistic reflection of the rarity of bankruptcy events.

This study distinguishes itself by moving beyond simple classification to a comparative analysis of model interpretability and performance. We evaluate three distinct modeling approaches: **Logistic Regression** (as a linear baseline), **XGBoost** (representing ensemble tree-based methods), and a **Deep Neural Network** (representing non-linear deep learning). Specifically, we assess their efficacy in predicting bankruptcy outcomes with a three-year horizon. Our findings reveal that while modern machine learning techniques like XGBoost offer superior recall in identifying at-risk firms, the "black-box" nature of

complex models necessitates a careful trade-off between predictive accuracy and interpretability. This paper provides empirical evidence supporting a progressive distress path, offering a practical framework for early warning systems.

2. Data & Method

2.1. Dataset Source and Sample Description

The empirical analysis is based on the **American Bankruptcy Dataset**, which provides a comprehensive panel of financial information for U.S. listed companies spanning the period from **1999 to 2018**. The dataset is structured on a firm-year basis, initially containing raw accounting data mapped to generic identifiers (X1 through X18).

After filtering for data consistency, the final sample consists of **78,682 firm-year observations** corresponding to **8,971 unique companies**. Each observation is labeled with a binary `status_label`:

- **0 (Alive/Operating)**: Represents healthy companies with normal financial operations.
- **1 (Failed/Bankrupt)**: Represents companies experiencing severe financial distress.

A defining characteristic of this dataset is its extreme class imbalance. The positive class (financial failure) constitutes approximately 6% of the total sample. This imbalance poses a significant challenge for predictive modeling and necessitates specific handling strategies (detailed in Section 2.4) to prevent model bias towards the majority class.

2.2. Feature Selection and Financial Indicators

To capture the multi-dimensional nature of corporate distress, we utilized **18 continuous financial indicators** as predictor variables. These features were renamed from their original codes to standard financial terminology to ensure interpretability. The feature set covers four critical dimensions of corporate health:

1. **Profitability & Efficiency**: e.g., *Net Income* (formerly X6), *Retained Earnings* (X2), *Gross Profit* (X16). These metrics assess the firm's ability to generate value from its resources.
2. **Liquidity & Solvency**: e.g., *Current Assets* (X1), *Total Liabilities* (X4), *Working Capital*. These measure the firm's capacity to meet short-term and long-term obligations.
3. **Leverage Structure**: e.g., *Long-term Debt* (X11), *Total Debt*. High leverage is often a precursor to default.
4. **Operational Scale**: e.g., *Total Assets* (X3), *Total Revenue* (X12), *Market Value* (X10).

2.3. Data Preprocessing Pipeline

Before feeding the data into machine learning algorithms, a rigorous preprocessing pipeline was applied to ensure data quality and model convergence.

- **Handling Missing Values and Encoding:** We first verified the integrity of the dataset. The categorical target variable `status_label` was encoded into a binary format (0/1).
- **Event-Time Construction for a Three-Year Horizon:** To align the predictive task with a fixed early-warning window, we do not classify all firm-years directly. Instead, for each firm that eventually fails, we identify its event year of bankruptcy (the last year in which it is labeled as failed or bankrupt) and construct an event-time index, defined as “`years_before_event = event_year - year`”. We then define the positive class as all firm-years with `years_before_event = 3`, i.e., observations taken three years before the failure event. For firms that never fail, we take their last available firm-year as a healthy observation and assign it to the negative class. This yields a focused prediction dataset in which each observation represents a company observed three years before either bankruptcy or continued survival.
- **Feature Scaling (Standardization):** Financial ratios often vary significantly in magnitude (e.g., *Total Assets* in billions vs. *Return on Equity* in decimals). To mitigate the impact of differing scales, particularly for gradient-based optimization in Neural Networks and Logistic Regression, we applied **StandardScaler** (Z-score normalization). This transformation centers each feature to a mean of 0 and a standard deviation of 1, ensuring that no single feature dominates the objective function due to raw magnitude.
- **Stratified Train-Test Split:** To evaluate model generalizability, the dataset was partitioned into a **Training Set (70%) and a Testing Set (30%)**. Crucially, the split was implemented with stratified sampling. This ensures that the proportion of failed and non-failed firms is preserved in both the training and testing subsets, providing a statistically representative evaluation environment.

2.4. Modeling Framework

We developed three distinct classification models to predict corporate distress, representing a progression from linear interpretability to non-linear complexity.

A. Logistic Regression (Baseline) We employed Logistic Regression as a baseline model. As a linear classifier, it provides high interpretability, allowing us to quantify the odds-ratio impact of each financial indicator. This model serves as a benchmark to determine whether complex non-linear relationships are necessary for this prediction task.

B. XGBoost (Extreme Gradient Boosting) To capture non-linear interactions between financial ratios (e.g., how high debt becomes risky only when combined with low liquidity), we implemented **XGBoost**. This ensemble tree-based algorithm is state-of-the-art for tabular data.

- *Handling Imbalance:* Instead of synthetic oversampling (SMOTE), we utilized the `scale_pos_weight` hyperparameter. This assigns a higher weight to the minority class (Failed) during gradient calculation, effectively penalizing the model more for missing a bankruptcy case than for misclassifying a healthy firm.

C. Deep Neural Network (DNN) We also experimented with a feed-forward neural network implemented in PyTorch to provide a deep learning benchmark for our tabular, highly imbalanced setting. Rather than relying on high-level Keras abstractions, we explicitly constructed a `nn.Module` class and

used `TensorDataset` and `DataLoader` objects to feed mini-batches of standardised financial features into the network. This design allows us to apply the same preprocessing (training-set standardisation) as in the linear model while leveraging GPU-friendly training and flexible handling of class weights in the loss function.

- **Architecture:** Architecture: The network consists of an input layer matching the dimensionality of the feature vector (18 financial variables), followed by three fully connected hidden layers with 64, 32 and 32 units respectively. Each hidden layer uses the Rectified Linear Unit (ReLU) activation function to introduce non-linearity:

1.Layer 1: `Linear(18, 64) + ReLU`

2.Layer 2: `Linear(64, 32) + ReLU`

3.Layer 3: `Linear(32, 32) + ReLU`

The final layer is a single-unit linear layer `Linear(32, 1)` that outputs a logit rather than a probability. During evaluation, we apply the Sigmoid function to this logit to obtain a predicted probability of failure for each firm-year.

- **Loss Function and Optimisation:** To handle the severe class imbalance in a principled way, we use the `BCEWithLogitsLoss` criterion with a positive-class weight (`pos_weight`) set to the ratio of negative to positive examples in the training data. This up-weights bankrupt firms in the loss function, ensuring that errors on the minority “failed” class receive much larger penalties than errors on the majority “alive” class. Model parameters are updated using the Adam optimiser with a learning rate of 1e-3, which provides stable convergence for this relatively small network without extensive manual tuning.
- **Compilation:** The model was compiled using the **Adam** optimizer for adaptive learning rates and **Binary Cross-Entropy** as the loss function.
- **Training Procedure:** The standardised training set is wrapped in a PyTorch `DataLoader` with a batch size of 128 and shuffled at each epoch; the test set is evaluated with a larger batch size of 256 and no shuffling. We train the network for 30 epochs, monitoring the average training loss after each epoch. In practice, the loss decreases only modestly—from roughly 1.26 around epoch 5 to about 1.18 by epoch 30—suggesting that the network has limited capacity to separate failing from healthy firms given the available signal and sample size. After training, we apply the Sigmoid function to the test-set logits and vary the decision threshold to compute precision, recall and F1 for the failed class, as reported in the Results section.

2.5. Evaluation Metrics

Given the imbalanced nature of the dataset, traditional accuracy is a misleading metric (a trivial model predicting "Alive" for everyone would achieve 94% accuracy). Therefore, our evaluation focuses on:

- **Recall (Sensitivity):** The proportion of actual bankruptcies correctly identified. This is the most critical metric for risk management.

- **Precision:** The proportion of predicted bankruptcies that were actual failures.
- **F1-Score:** The harmonic mean of Precision and Recall.

3. Results & Discussion

3.1. The Divergence of Financial Trajectories (Exploratory Analysis)

Prior to modeling, a descriptive analysis of the dataset reveals significant structural differences between healthy and distressed firms as early as three years before the actual failure event.

- **Profitability Gap:** Firms approaching bankruptcy exhibit a marked deterioration in profitability. Our mean-difference analysis shows that profitability collapses well before the bankruptcy event. On average, failing firms have much lower Net Income and Retained Earnings than healthy firms, with relative mean differences exceeding -100% . This implies a substantial transition from modest profits to large cumulative losses in the three years leading up to bankruptcy, consistent with the idea of a prolonged distress phase rather than a sudden collapse.
- **Liquidity Crunch:** Distressed firms show a significantly higher ratio of *Total Debt* relative to their *Total Assets*, indicating a reliance on leverage to sustain operations as internal cash flow dries up.
- **Scale Contraction:** Interestingly, the data suggest a "shrinking scale" phenomenon. Failing firms tend to have lower *Market Value* and *Total Assets* compared to their surviving peers, reflecting the market's early pricing of their distress risk. These statistical disparities confirm our hypothesis: **bankruptcy is not a sudden jump but a progressive deterioration trajectory that is statistically detectable at t-3.**

3.2. Comparative Analysis of Model Performance

We evaluated the three models (Logistic Regression, XGBoost, DNN) based on their ability to classify firms 3 years in advance. Given the severe class imbalance (approx. 6% failure rate), we prioritize Recall (Sensitivity) and F1-Score over simple Accuracy.

(1) The Baseline: Logistic Regression. Logistic Regression provided a robust and interpretable baseline. It successfully identified the main linear drivers of bankruptcy, such as low profitability and high leverage. However, once we up-weighted the minority class using `class_weight="balanced"`, the model achieved high recall on failing firms (around 0.78) at the cost of extremely low precision (around 0.06). In other words, the logistic model must flag a very large set of firms as "at risk" in order to catch most future bankruptcies. This pattern suggests that purely linear decision boundaries are not flexible enough to capture the complex, interactive distress signals in the data—for example, firms that are currently profitable but combine fragile liquidity with dangerous short-term debt structures.

(2) The Champion: XGBoost emerged as the most useful model for practical early-warning. By utilising the `scale_pos_weight` parameter to penalise false negatives, the model achieves high overall accuracy (about 0.82 on the test set) while still recovering roughly 30% of the failing firms three years before bankruptcy and maintaining a substantially higher precision (around 0.11) than the logistic

baseline. This indicates that tree-based ensembles are particularly effective at handling the non-linear “tipping points” in financial ratios. In our setting, XGBoost offers the most attractive trade-off: it meaningfully reduces the risk of missing a bankruptcy while keeping the false-alarm rate at a level that is still manageable for practitioners.

(3) The Limitation of Deep Learning (DNN) Contrary to the general trend in unstructured data (like images or text), the Deep Neural Network **underperformed** in this tabular setting. The DNN struggled to converge on the minority class, often biasing its predictions towards the majority ("Alive") class.

- *Interpretation:* We attribute this to the "Vanishing Gradient" problem inherent in highly imbalanced datasets where the signal from the minority class is overwhelmed by the majority noise during backpropagation. This finding reinforces a critical insight: **Complex architectures do not automatically yield better results for tabular financial data; feature engineering and ensemble methods often prevail.**

3.3. Key Determinants of Corporate Failure

Analyzing the *Feature Importance* from the XGBoost model allows us to rank the predictors of distress. The results are consistent with financial theory:

1. **Retained Earnings & Net Income:** These are the top predictors. A consistent inability to generate internal capital is the single strongest warning sign three years out.
2. **Market Value:** The equity market often detects trouble before the accounting statements do. A plummeting market capitalization acts as a leading indicator.
3. **Total Liabilities:** High absolute debt levels, especially when unmatched by asset growth, serve as a critical risk factor.

3.4. Discussion: The Trade-off Between "Black Box" and Interpretability

Our results highlight a practical dilemma for financial stakeholders.

- **Logistic Regression** offers full transparency (we know exactly how much probability increases for every \$1 of debt) but misses subtle risk patterns.
- **XGBoost provides substantially higher predictive accuracy than the logistic baseline and a more balanced combination of recall and precision on the failed class, but it operates as a “black box.”** For a three-year early-warning system, recall on failing firms remains crucial—it is generally more costly for a bank to lend to a future bankrupt firm (a default loss) than to reject a healthy one (an opportunity cost). At the same time, a screening tool that flags almost every firm as distressed is not operationally useful. From this practical perspective, XGBoost strikes the most appealing compromise between missed defaults and false alarms and is therefore our recommended model for deployment in a 3-Year Early Warning System.

4. Conclusion

4.1. Summary of Findings

This study set out to investigate whether corporate bankruptcy is predictable three years prior to the event, utilizing a dataset of approximately 8,900 U.S. firms. By benchmarking Logistic Regression, XGBoost, and Deep Neural Networks, we provide empirical evidence for the "progressive deterioration" hypothesis. Our analysis yields three primary conclusions:

- **Predictability at t-3:** Corporate distress is not an instantaneous shock. Distinct financial patterns—specifically in **Retained Earnings** and **Market Value**—diverge significantly between healthy and failing firms fully three years before default.
- **Model Superiority:** XGBoost proved to be the most effective algorithm for this task. It successfully handled the severe class imbalance and non-linear feature interactions, outperforming the linear baseline (Logistic Regression) in Recall and F1-score.
- **The "Deep Learning" Trap:** Contrary to expectations, the Deep Neural Network (DNN) underperformed on this tabular dataset. This suggests that for structured financial data with limited sample size and high imbalance, complex architectures do not necessarily yield better generalization than robust ensemble methods.

4.2. Practical Implications

The results have immediate relevance for financial stakeholders:

- **For Creditors:** The dominance of "Retained Earnings" as a predictor suggests that cumulative profitability is a more reliable long-term signal than short-term liquidity ratios. A "3-Year Early Warning System" based on the XGBoost framework could allow banks to intervene or restructure debt well before a crisis becomes terminal.
- **For Investors:** The high feature importance of "Market Value" indicates that equity markets often price in distress risk before it fully materializes in accounting statements. This confirms the efficiency of market signals as a leading indicator.

4.3. Limitations

While robust, this study is subject to specific limitations:

- **Geographical and Temporal Scope:** The dataset is restricted to U.S. firms (1999–2018). The model's applicability to emerging markets, where accounting standards and bankruptcy laws differ, remains to be tested.
- **Feature Constraints:** Our model relies exclusively on quantitative financial ratios. It does not incorporate qualitative factors (e.g., management changes, audit opinions) or macroeconomic indicators (e.g., interest rate cycles), which likely influence bankruptcy risk.
- **Static vs. Dynamic:** We treated the "3-year prior" prediction as a static classification problem. This approach ignores the *rate of change* in financial variables year-over-year, which a time-series model might capture more effectively.

4.4. Future Directions

Future research should aim to integrate **unstructured data** (such as textual analysis of annual reports or news sentiment) to capture "soft" signals of distress. Additionally, moving from static classifiers to

Sequence Models (e.g., LSTM or RNNs) would allow researchers to model the *trajectory* of financial decay rather than just a snapshot in time.