Vienna University of Economics and Business

Master Thesis Data Methodology

The Agony and the Ecstasy: Constructing a "Crash-Filtered" Equity Index using Machine Learning

Author:
Tristan Leiter

Submission Date:
December 16, 2025

Supervisor:
Univ.Prof. Dr. Kurt Hornik
Department of Finance, Accounting and Statistics

Chapter 1

Data

1.1 Constituent Universe

The study utilizes the Center for Research in Security Prices (CRSP) database, specifically the *Stock Security Information History* (`stksecurityinfohist`) and *Monthly Stock File* (`msf`) accessed via WRDS. To ensure data quality and constituent continuity, the investable universe is constructed through a four-step process:

1. **Whole Sample:** The initial sample includes all unique permanent identifiers (`permno`) in the CRSP database, covering the entire history of US-listed securities to eliminate survivorship bias.
2. **Aggregation:** Data is aggregated by `permno` to define the distinct trading lifetime, calculated as the duration between the earliest listing date and the final removal date.
3. **Filtering Criteria:** The universe is subjected to the following filters:
   - **Security Type:** Restricted to common equities (`securitytype="EQTY"`, `sharetype="COM"`), explicitly excluding ETFs and other Mutual Funds.
   - **Exchange:** Limited to major US exchanges (NYSE, AMEX, NASDAQ).
   - **Historical Depth:** Excludes securities removed prior to January 1, 1966.
   - **Minimum Lifetime:** Requires a listing duration of at least 5 years to learn on multiple datapoints from each firm and to ensure the same methodology for the computation of the dependent variable.

1.1.1 Sample Overview

The final dataset spans from January 1960 to December 2024, comprising 1,064,324 firm-month observations across 3,629 unique equity issuers (Table 1.1).

Table 1.1: Aggregate Sample Statistics

Statistic	Value
Total Unique Firms (N)	3,629
Total Firm-Month Observations	1,064,324
Average Months per Firm	293
Sample Period	1960-01-31 to 2024-12-31

1.1.2 Industry Classification

Economic activity is categorized using the **North American Industry Classification System (NAICS)**. NAICS codes are available in CRSP from August 2001; for securities delisted prior to this, legacy SIC codes are mapped to equivalent NAICS sectors to ensure complete coverage. These classifications will later on by changed to match the ones by the Global Industry Classification Standard (GICS).

Table 1.2: Constituent Distribution by Industry Sector (NAICS/SIC Hybrid)

Industry Sector	Count	Percentage (%)	Source
Manufacturing	1285	35.4	NAICS
Finance and Insurance	547	15.1	NAICS
Professional, Scientific, & Technical Services	237	6.5	NAICS
Real Estate and Rental and Leasing	217	6.0	NAICS
Information	215	5.9	NAICS
Mining, Quarrying, and Oil & Gas Extraction	188	5.2	NAICS
Admin. Support & Waste Mgmt	139	3.8	NAICS
Retail Trade	131	3.6	NAICS
Wholesale Trade	128	3.5	NAICS
Management of Companies and Enterprises	108	3.0	NAICS
Transportation and Warehousing	102	2.8	NAICS
Utilities	87	2.4	NAICS
Accommodation and Food Services	67	1.8	NAICS
Health Care and Social Assistance	57	1.6	NAICS
Construction	55	1.5	NAICS
Educational Services	18	0.5	NAICS
Arts, Entertainment, and Recreation	16	0.4	NAICS
Agriculture, Forestry, Fishing & Hunting	15	0.4	NAICS
Other Services (except Public Admin)	14	0.4	NAICS
Public Administration	2	0.1	NAICS

Chapter 2

Methodology

2.1 Classification of a Catastrophic Stock Implosion (CSI)

This study adopts a rule-based detection algorithm to identify "Catastrophic Stock Implosions" (CSI), distinguishing permanent capital destruction from standard volatility. Following Tewari et al. (2024), a stock is classified as imploded if it satisfies a specific sequence of drawdown and non-recovery, governed by three parameters:

1. **Initial Crash Threshold ($C = -0.8$):** A drop of at least 80% from the rolling all-time high.
2. **Zombie Period ($T = 18$ months):** A post-crash monitoring window to confirm the permanence of the loss.
3. **Recovery Ceiling ($M = -0.2$):** The price must remain below 80% of the pre-crash peak throughout the entire Zombie Period.

2.1.1 Detection Algorithm

For stock i at month t, let $P_{i,t}^{max} = \max(P_{i,0}, ..., P_{i,t})$ be the high-water mark. A candidate implosion is flagged if the drawdown $D_{i,t} \leq C$. The event is confirmed as a CSI if and only if:

$$\max(P_{i,t+1}, ..., P_{i,t+T}) \leq P_{i,t}^{max} \times (1 + M)$$

This confirms the stock has become "dead capital." The "Implosion Date" is recorded as the first month t where these conditions are met.

2.1.2 Temporal and Sectoral Distribution of Implosions

The algorithm identified 2,351 distinct failure events. Figure 2.1 illustrates the temporal clustering of these events, showing distinct spikes corresponding to the Dot-Com Bubble (2000-2002) and the post-COVID correction (2021-2022).

Figure 2.3 further decomposes these events by industry, highlighting how different sectors drive the failure rate in different market cycles (e.g., Tech in 2000 vs. Biotech/Pharma in 2021).

2.2 Descriptive Statistics of the Modeling Universe

To enable predictive modeling, the data was transformed into a target-oriented panel. The primary target variable, `TARGET_12M`, is set to 1 if the stock is within 12 months of a confirmed CSI event.

2.2.1 Post-Implosion Outcome Analysis

We analyze the long-term fate of firms after the 18-month "Zombie" period to quantify the "Agony and Ecstasy" dynamic. Figure 2.4 visualizes a random sample of these trajectories, with the blue shaded region representing the 18-month stagnation period.

As shown in Table 2.1, while 30% of firms stagnate or delist ("Agony"), approximately 33% achieve a strong recovery ("Ecstasy"; defined as a recovery with a subsequent geometric annualized return of more than 10%), justifying the need for a dynamic exclusion model rather than a permanent blacklist.

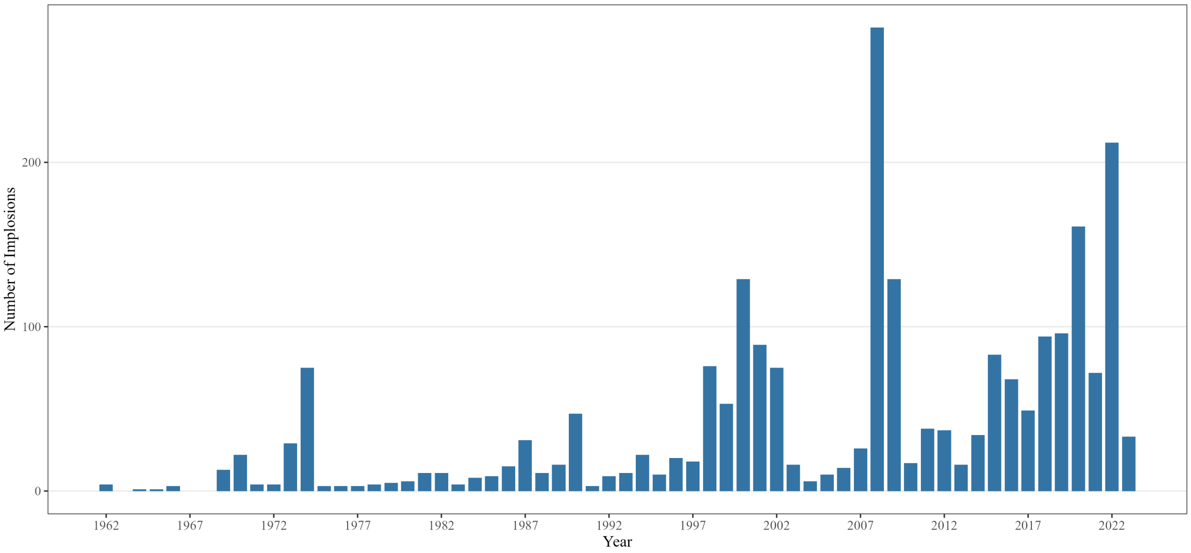
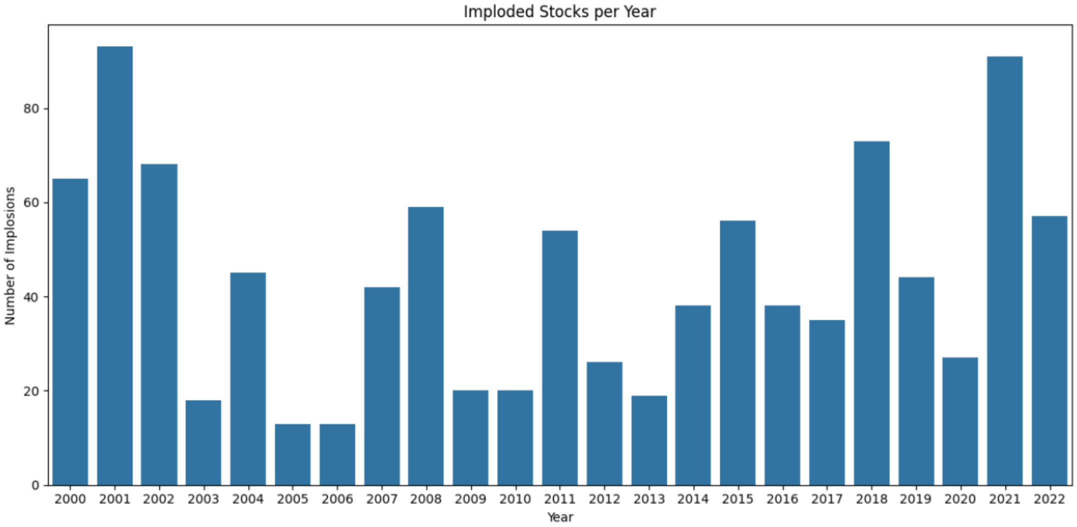


Figure 2.1: **Temporal Distribution of Catastrophic Implosions (1960–2024).** The bar chart displays the frequency of identified CSI events by year. Significant clustering is observed during major market corrections.



(a) Number of Implosions per Year

Figure 2.2: **Temporal Distribution of Catastrophic Implosions (1960–2024).** The bar chart is obtained from Tewari et al. (2024) and it displays the frequency of identified CSI events by year. Significant clustering is observed during major market corrections.

2.3 Class Imbalance

An interesting insight is that the number of US public companies in this constituent universe grows monotonically over time. Typically, the number of US public companies peaked in 1996 (at roughly 8,000) and then declined steadily to about 4,000 by 2012 (the "Listing Gap").

This is an effect of the initial filter (lifetime years greater than 4.99), which strongly reshapes the universe. I filtered out the penny stocks and short-lived IPOs, which might distort the ML-models. Thus, with this approach the model would not be appicable applicable to IPOs or young companies (age less than 5).

The "Look-ahead" Bias: In the 1960s/70s, the bars are very low. This is likely because only the "winners" from that era had enough data to survive the filters and make it into digital databases like WRDS in a clean format.

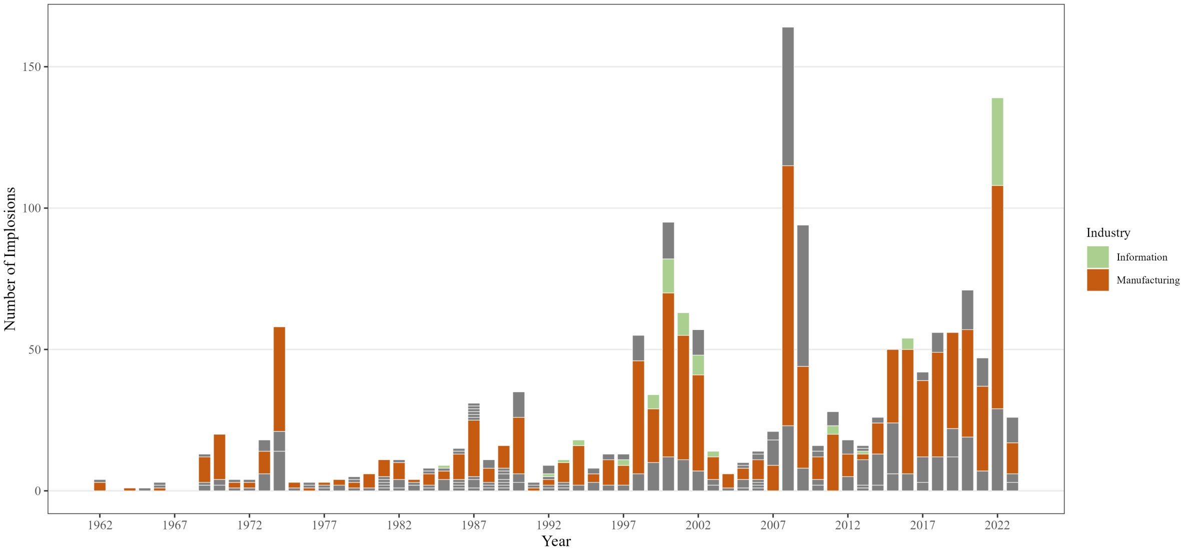


Figure 2.3: **Top 3 Imploding Industries per Year.** Stacked bar chart showing the three sectors contributing the most failures in each calendar year. Colors represent distinct NAICS sectors.

Table 2.1: Post-Implosion Outcomes

Outcome Category	Count	Percentage
Recovery (Low Growth)	866	36.84%
Recovery (Strong Growth)	768	32.67%
Implosion (Stagnation)	708	30.11%
Zombie State (Delisted)	9	0.38%

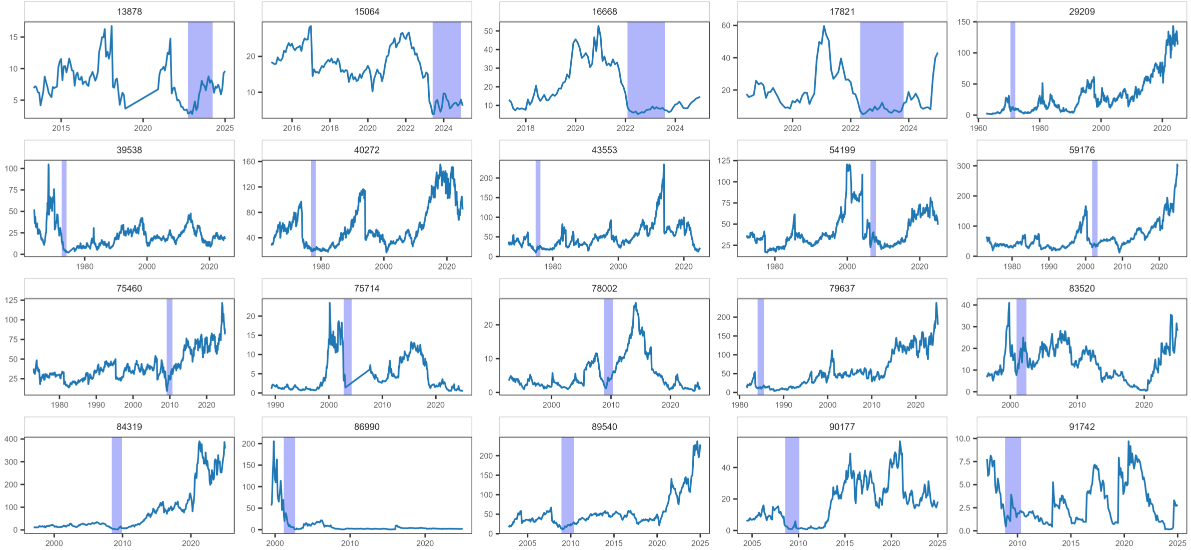


Figure 2.4: **Sample of Catastrophic Implosion Trajectories.** Price history for 20 randomly selected failed firms. The shaded blue area indicates the 18-month "Zombie Period" (T) following the initial crash, during which the price failed to recover above the threshold (M).

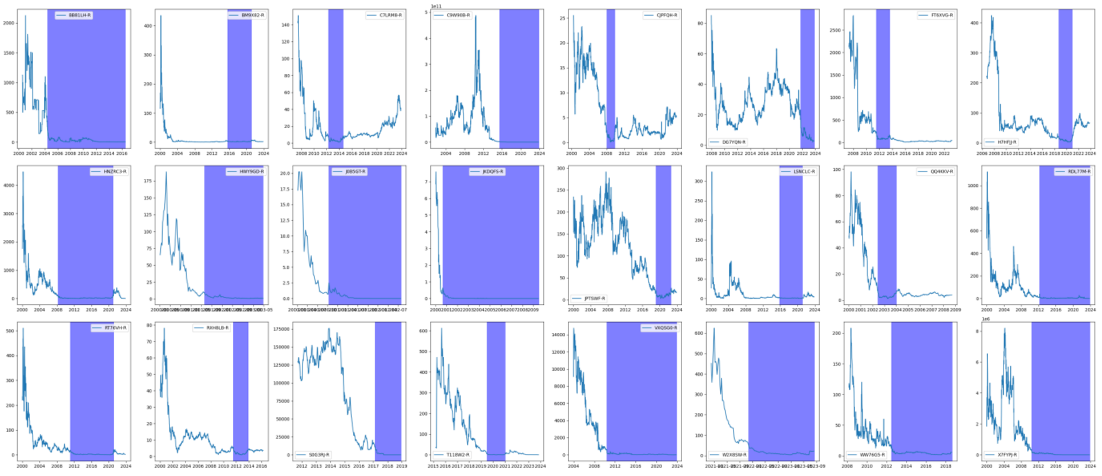


Figure 2: Sample of Implosions

Figure 2.5: **Sample of Catastrophic Implosion Trajectories.** The chart is obtained from Tewari et al. (2024). The shaded blue area indicates the 18-month "Zombie Period" (T) following the initial crash, during which the price failed to recover above the threshold (M).

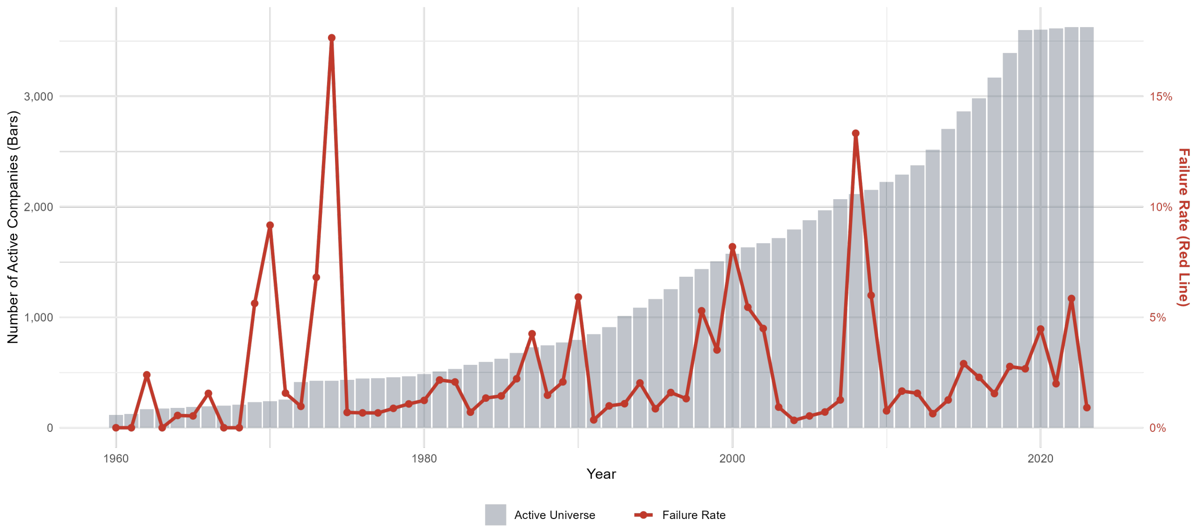


Figure 2.6: **CSI vs Market Breadth.** The grey bar charts show the number of companies (constituent universe) per year (left hand side). The red line depicts the failure rate (=CSI event) per year. It spikes during the crashes of 1973-74 (Oil Shock), 2000-2001 (DotCom) and GFC (2008). The Covid-Crisis (2020) is not long enough for a lot of companies to match the CSI definition (they recover quickly).

Chapter 3

Questions

3.1 Is the CSI-methodology sound?

I am critically evaluating the soundness of the Catastrophic Stock Implosion (CSI) classification as defined in the literature. My preliminary analysis of the WRDS dataset reveals that approximately 33% of firms classified as "imploded" eventually stage a strong recovery ("Ecstasy" outcome).

Question: Do we want to strictly predict *terminal* catastrophic implosions (no recovery allowed), or do we accept the broader definition of "severe distress"?

- *Option A:* Filter the target $Y = 1$ to only include firms that eventually delist or never recover. This lowers the sample size but purifies the "Catastrophic" label.
- *Option B (Proposed):* Maintain the broader definition (Drawdown $\leq -80\%$) because predicting deep distress is valuable regardless of the final outcome. We can potentially add a secondary model later to estimate recovery probability.

3.2 Which frequency to use?

Tewari et al. (2024) utilize a static annual framework (t predicts $t + 1$). However, accounting data is available quarterly, and market data is available monthly.

Question: Do you agree with shifting to a *quarterly rolling window* approach?

- Instead of one prediction per year, we update every quarter using the latest available fundamental and market data.
- *Target Definition:* $Y_{i,t} = 1$ if stock i implodes within the next 4 quarters $[t, t + 4]$.
- *Rationale:* This better mimics active risk management and significantly increases the effective sample size ($N \times T$), though it introduces serial correlation that must be addressed during validation.

3.3 Handling "Zombie" Periods (Censoring)

The reference paper explicitly removes data points following an implosion from the training set. They argue that investors are interested in avoiding the crash, not modeling the subsequent stagnation.

Question: Is this censoring methodologically correct for our goals?

- My understanding is that we are building an *Ex-Ante* Early Warning System (predicting entry into distress), not a survival analysis model.
- Therefore, removing "Zombie" rows prevents look-ahead bias and forces the model to learn only pre-crash signals.
- Including data after an "Implosion" could allow our model to reduce the impact of False Positives. This could be dealt with using a rolling window approach.

3.4 Probabilistic Risk Score vs. Binary Classification

I think the reasonable way would be to go for the classic Binary Classification case. If a firm fulfills the CSI characteristics, it is classified as a 1 and 0 otherwise. Possible extensions could include the following:

While the target variable is binary (Implosion vs. Survival), classification models (e.g., Logistic Regression, XGBoost) natively output a conditional probability $P(Y = 1|X)$.

Question: Can we frame the final output as a continuous "Implosion Probability Score" rather than a hard binary flag?

- *Context:* While the objective function is classification (e.g., `binary:logistic`), treating the output as a continuous probability allows for ranking stocks by risk (e.g., "Top Decile Risk") rather than treating all flagged stocks equally.
- This effectively bridges the gap between classification and regression, providing a more granular tool for portfolio construction without changing the underlying binary nature of the ground truth.

3.5 The Precision-Recall Trade-off

Tewari et al. (2024) prioritize Recall (catching as many crashes as possible) over Precision, arguing that missing a crash is worse than a false alarm. However, in a practical investment universe, low precision leads to excessive exclusion of healthy stocks.

Question: For the purpose of this thesis, should I optimize for a balanced metric (like ROC-AUC) or specifically target a high-Recall metric (like F2-Score)? For Tewari et al. (2024) it seems to work out as False Positives are less than 3%.

3.6 Proposed Modeling Framework

3.6.1 Prediction Horizon and Frequency

In contrast to the static annual approach used in prior literature (Tewari et al., 2024), we implement a quarterly rolling window framework. At each quarter t, the model utilizes the most recent available information (market data from month t and fundamental data with appropriate reporting lags) to predict the probability of a Catastrophic Stock Implosion (CSI) occurring within the subsequent 4 quarters $[t + 1, t + 4]$.The target variable is defined as:

$$Y_{i,t} = \begin{cases} 1 & \text{if stock } i \text{ triggers a CSI event in } [t+1, t+4] \\ 0 & \text{otherwise} \end{cases}$$

This increases the effective sample size and allows for "point-in-time" risk assessment, enabling the detection of distress signals as they emerge rather than waiting for fiscal year-end updates.

3.6.2 Censoring and Bias Prevention

To ensure the model functions as an *ex-ante* early warning system, we apply strict censoring to the "Zombie Periods." Data points corresponding to the 18-month monitoring window following a triggered implosion are excluded from the training set. This prevents the model from "learning" obvious distress signals (e.g., price already down 90%) after the event has occurred, forcing it to focus exclusively on leading indicators present *before* the crash.

Chapter 4

Refined Research Design & Evaluation

Based on the preliminary diagnostics of the failure rate (which correctly captures the 1974, 2000, and 2008 crises) and the constituent universe (which is biased toward mature firms), the following methodological refinements are proposed to ensure the model captures structural distress rather than random market noise.

4.1 Cross-Validation Strategy: The Expanding Window

Unlike standard time-series forecasting, corporate implosions are rare events that cluster heavily during specific economic regimes (e.g., the Dot-Com Bubble or the Great Financial Crisis). A standard rolling window (e.g., "train on 10 years, shift forward 1 year") risks "forgetting" these critical lessons. For instance, a model trained on a 2010–2019 window would learn that "stocks never fail," leading to catastrophic under-prediction during a subsequent shock like COVID-19.

To mitigate this, we adopt an **Anchored Expanding Window** (Walk-Forward) approach:

1. **Anchored Start:** The training set always begins at the start of the sample (1960).
2. **Expansion:** For each iteration k, the training window extends from 1960 to year t_k.
3. **Out-of-Sample Test:** The model is tested on the subsequent year $t_k + 1$.

This ensures that once the model "sees" a crisis (e.g., 2008), that data remains in its memory banks forever, preventing the model from becoming complacent during long bull markets.

4.2 Feature Engineering: Common-Sizing and Standardization

To ensure the model learns fundamental fragility rather than nominal size, all balance sheet features are standardized prior to training. The preprocessing pipeline follows three steps:

1. **Common-Sizing:** All nominal values are scaled by Total Assets (AT) or Market Capitalization to create ratios (e.g., Debt/Assets, Cash/Assets). This neutralizes the size factor, ensuring a small cap firm with 90% leverage is treated similarly to a large cap firm with 90% leverage.
2. **Winsorization:** Financial ratios frequently exhibit extreme outliers (e.g., a firm with near-zero equity creating infinite ROE). We clip all continuous features at the 1st and 99th percentiles to prevent these outliers from distorting the gradient descent process.
3. **Z-Score Normalization:** Finally, features are centered ($Mean = 0$) and scaled ($SD = 1$) to ensure convergence stability across different feature types.

4.3 Evaluation Philosophy: Fragility Identification over Market Timing

Given that the exact timing of an implosion is often triggered by exogenous shocks (e.g., a pandemic or oil embargo) which are inherently unpredictable, the objective of this thesis is refined from "Predicting Crashes" to **"identifying Fundamental Fragility."**

The model is evaluated not on its ability to predict the exact day of a crash, but on its ability to sort firms into "Risk Buckets" that disproportionately capture failures when a shock occurs.

4.3.1 The "Sorting" Test (Quantile Portfolios)

To validate this, we utilize a **Quantile Sorting Approach** rather than simple binary accuracy:

1. At the start of each test year, the model ranks the entire universe of stocks based on their predicted "Implosion Probability" (P).
2. Stocks are sorted into Quintiles ($Q1$ = Safest, $Q5$ = Most Fragile).
3. We track the realized failure rate within the top quintile ($Q5$) over the subsequent 12 months.

Figure 4.1: **Conceptual Validation: The "Dry Forest" Hypothesis.** If the model is successful, the High-Risk portfolio (Q5) should capture the majority of realized failures during crisis years, confirming that the model correctly identified the "driest wood" in the forest before the spark hit.

4.3.2 Success Metrics

Success is defined by the **Top-Decile Lift**:

$$\text{Lift} = \frac{\text{\% of Failures captured in Top Decile}}{\text{10\% (Random Guess Baseline)}}$$

A Lift > 1 indicates the model is adding information. In financial distress literature, a Lift > 3 (capturing 30% of failures in the top 10% of sorted stocks) is considered a robust result.

Bibliography

Zaki Tewari, Michal Galas, and Philip Treleaven. Predicting catastrophic stock implosions: A machine learning approach. Technical report, University College London, 2024. Working Paper.