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 14, 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. **Selected Attributes:** A "Terminal Attribute" approach is adopted, selecting descriptive characteristics (e.g., Ticker, Exchange) from the security's final available record to ensure consistent grouping.
4. **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 **?**, 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 2.2 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.

THIS WILL BE CRUCIAL. HOW SHOULD IT BE CLASSIFIED?

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 2.3 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 strong recovery ("Ecstasy"), justifying the need for a dynamic exclusion model rather than a permanent blacklist.

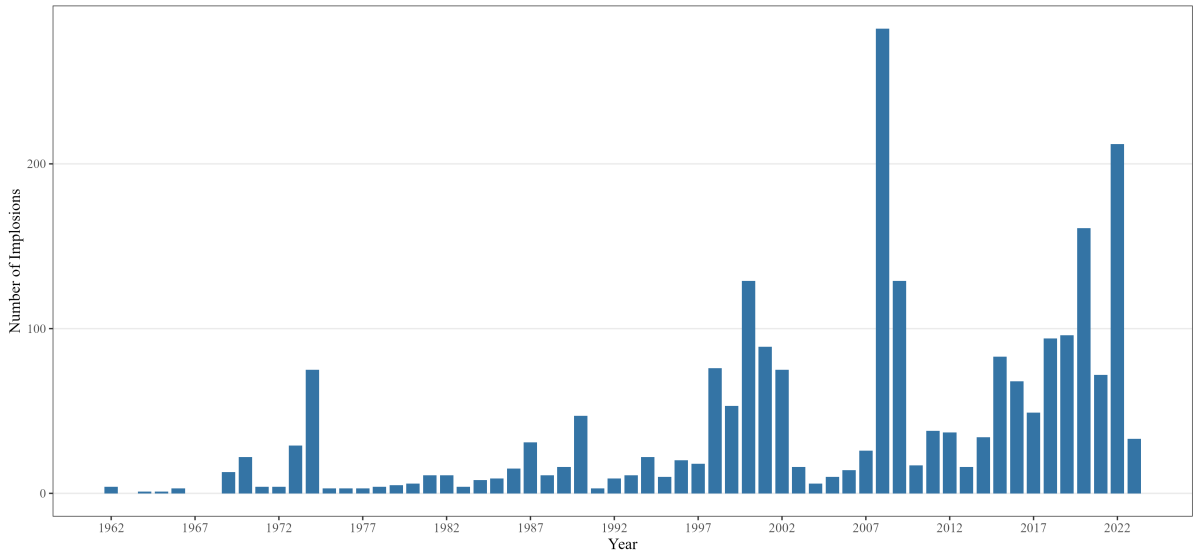

Figure 2.1: **Temporal Distribution of Catastrophic Implosions (1960–2024).** The bar chart displays the frequency of identified CSI events by year. Significant clustering is observed during major market corrections.

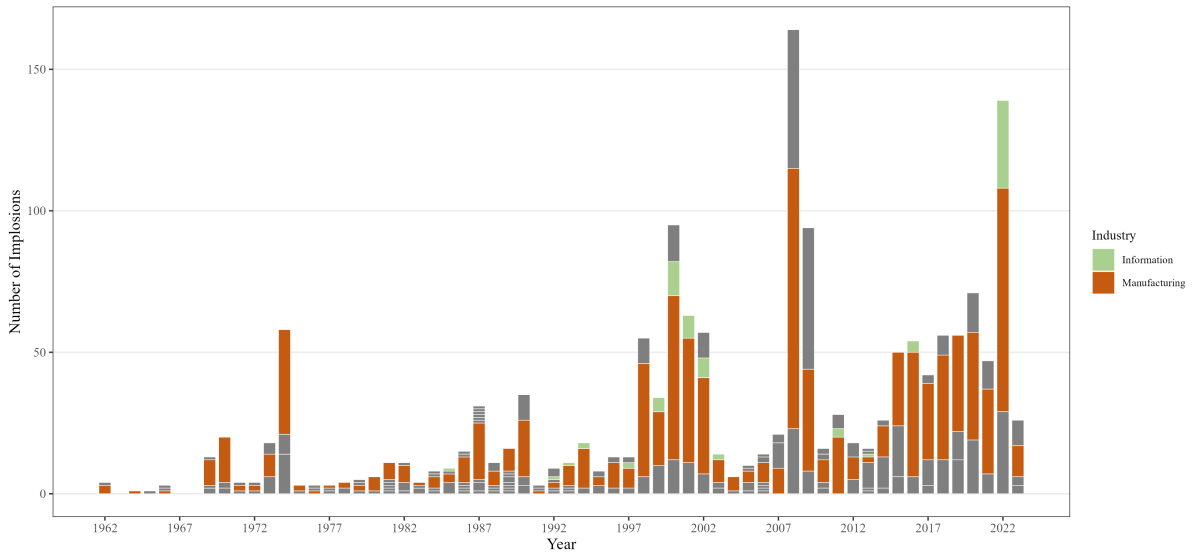

Figure 2.2: **Top 3 Imploding Industries per Year.** Stacked bar chart showing the three sectors contributing the most failures in each calendar year. Colors represent distinct NAICS sectors.

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%

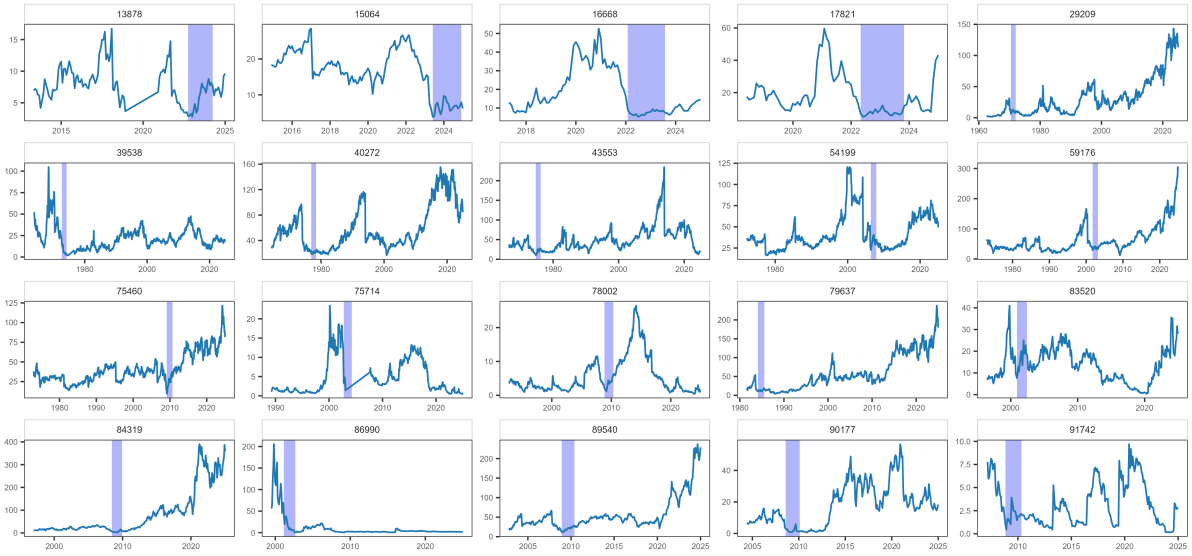

Figure 2.3: **Sample of Catastrophic Implosion Trajectories.** Price history for 20 randomly selected failed firms. The shaded blue area indicates the 18-month "Zombie Period" (T) following the initial crash, during which the price failed to recover above the threshold (M).

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 daily.

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

- Instead of one prediction per year, we update $P(\text{Implosion})$ every month using the latest available fundamental and market data.
- *Target Definition:* $Y_{i,t} = 1$ if stock i implodes within the next 12 months $[t, t + 12]$.
- *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.

3.4 Probabilistic Risk Score vs. Binary Classification

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)?

3.6 Proposed Modeling Framework

Based on the characteristics of the dataset and the research objectives, this study adopts a **probabilistic classification framework** tailored for active risk management. The proposed methodology deviates from static annual predictions in favor of a dynamic, rolling-window approach that better mimics real-world investment constraints.

3.6.1 Prediction Horizon and Frequency

In contrast to the static annual approach used in prior literature (**?**), we implement a monthly rolling window framework. At each month 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 12 months $[t + 1, t + 12]$.The target variable is defined as:

$$Y_{i,t} = \begin{cases} 1 & \text{if stock } i \text{ triggers a CSI event in } [t+1, t+12] \\ 0 & \text{otherwise} \end{cases}$$

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.

3.6.3 Probabilistic Risk Scoring

While the ground truth is binary (Implosion vs. Survival), the model is optimized using a Log-Loss (Cross-Entropy) objective function to output a continuous conditional probability score:

$$\hat{P}_{i,t} = P(Y_{i,t} = 1 \mid \mathbf{X}_{i,t})$$

This output is interpreted as a "Implosion Risk Score" ranging from 0 to 1. This granular metric allows for the ranking of securities by risk level (e.g., separating "High Risk" from "Critical Risk") and enables the construction of flexible exclusion thresholds (e.g., "Exclude top decile") rather than relying on a rigid binary classification boundary.