The Agony and the Ecstasy

Constructing a "Crash-Filtered" Equity Index using Machine Learning

Tristan Leiter

Vienna University of Business and Economics

January 28, 2026

The "Agony and Ecstasy" of Indexing

- **The Passive Investing Dilemma:** Indices capture the "Ecstasy" of extreme winners but systematically force investors to hold the "Agony" of imploding stocks.
    - Cembalest (2014) finds that equity indices are driven by a small tail of winners, while approx. 40% of constituents suffer "catastrophic declines" ($> 70\%$ drawdown) without recovery.
- Tewari et al. (2024) formalize these events as **Catastrophic Stock Implosion (CSI)**.
    - It is a distinct event: a severe price downturn followed by a "zombie" period of prolonged stagnation.
- Passive investing captures the winners, but fails to filter the "Agony" until it is too late. Standard metrics fail to distinguish between **recoverable volatility** and **terminal implosion**.

Research Question

Main Research Question:

To what extent does a 'Crash-Filtered' equity index, constructed via probabilistic implosion modeling, generate superior risk-adjusted returns compared to the market benchmark and traditional minimum-volatility strategies?

Subquestions:

- How does the integration of Autoencoders for feature engineering impact the Average Precision of Ensemble models compared to those trained solely on raw financial data?
- To what extent do Ensemble methods reduce the False Positive Rate (classifying recoverable volatility as CSI) compared to traditional volatility-based exclusion strategies, while maintaining Recall?
- Which features are most important for distinguishing between 'Zombie' firms (CSI) and non-zombie firms?

Limits of Traditional Models

- **The "Quality Trap":** Perceived safety signals can be misleading. Penman and Reggiani (2018) suggest that low B/P ratios often reflect uncertainty rather than value, while profitability measures lose predictive power over long horizons.
- **The False Positive Dilemma:** Traditional bankruptcy models (e.g., Altman (1968)) and risk scaling strategies, like Minimum-Volatility (MinVol.), fail to distinguish between 'good' volatility (growth) and 'bad' volatility (implosion), systematically excluding winners.

Methodology I: Dependent Variable

Goal: Set up the dependent variable Catastrophic Stock Implosion (CSI)
Following Tewari et al. (2024), a stock is classified as a CSI ($y = 1$) if it satisfies:

- **Initial Crash (C):** $> 80\%$ drawdown from trailing peak (beginning of the "zombie" period).
- **Non-Recovery:** A maximum cumulative return of -20% in the "zombie" period.
- **Zombie Period:** Duration of 1.5 years.

Tewari et al. (2024) employ a yearly prediction horizon ($h = 12$ months), which this thesis follows:

$$y_{i,t} = \begin{cases} 1 & \text{if stock } i \text{ triggers } C = 80\% \text{ within } [t, t+h] \text{ and zombie criteria met,} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Methodology II: Modelling

- Using **Autoencoders** on the feature engineered data to test their utility for either feature denoising or signal generation.
- **Supervised Learning:** Training ensemble methods (Random Forest, XGBoost, CatBoost, LightGBM) on both raw data and latent features to predict the probability of CSI.

Cross-Validation optimization will be conducted via **Average Precision (AP)**. Since the dataset is imbalanced, AP provides a more robust signal for hyperparameter tuning without committing to a specific decision threshold.

The models will be evaluated based on **Recall at fixed FPR**.
A constraint-based metric aligns more closely with the practical "Risk-Budget" in Portfolio Management. The objective is to maximize the number of Agony stocks identified (Recall) whilst capping the exclusion of Ecstasy stocks.

Methodology III: Index Construction

To ensure a proper **Out-of-Sample test**, the dataset will be split into three parts:

- Training set for Cross-Validation
- Test set for Model-selection
- Out-of-Sample set for backtesting

The best-performing model within the test-set will be used to predict a CSI for each firm in the consecutive year. All firms exceeding the probability threshold, θ, will be removed. In this context, θ is selected based on the desired FPR rate (for example 3% or 5%). Additionally, the "Crash-Filtered" Index will be rebalanced annually at the end of each calendar year.

Index Construction

The "Crash-Filtered" Index systematically excludes constituents where the predicted probability $\hat{p}_{CSI} > \theta$. Unlike standard classifiers that use a default $\theta = 0.5$, this threshold is **dynamically calibrated** to satisfy a specific risk constraint (e.g., $FPR \leq 5\%$), ensuring the exclusion rate aligns with the investor's "budget" for opportunity cost.

CRSP

The Universe (CRSP)

- **Scope:** US Common Equities (NYSE, AMEX, NASDAQ).
- **Timeline:** 1998 - 2024.
- **Size:** $3,263$ unique firms, 51,773 firm-year observations.
- **Constraint:** Minimum listing lifetime of 5 years to ensure sufficient learning history (13 years on average per firm).
- **CSI-Events:** $2,236$ CSIs in Total.

Category	Cohort Distribution	
	Imploded Firms	Never Imploded
High Growth ($> 10\%$)	4.9%	40.5%
Moderate Growth ($5 - 10\%$)	9.1%	25.9%
Low Growth ($2 - 5\%$)	8.2%	9.0%
Stagnation ($-2 - 2\%$)	0.0%	0.1%
Value Destruction (temp. Recovery)	64.8%	15.4%
Value Destruction (No Recovery)	2.2%	0.8%
Unknown	10.8%	8.4%

Note: Categories are measured by the average geometric return.

Table 1: Categorization of CSI events dependent on event type.

Compustat

Features

- **Accounting:** Balance-Sheet and Income-Statement variables.
- **Macro:** Variables with possible interactions with accounting variables (interest rate, others), obtained from the FRED database.
- **Other:** Specifically targeting "Zombie" precursors:
  - *Employees* (Number of employees).
  - *Rental expenses*

Compustat

Category	**Variable Code**	**Description**
Balance Sheet: Assets	`at` / `act`	Total Assets / Total Current Assets
	`che` / `ivst`	Cash & Short-Term Inv. / Short-Term Investments
	`rect` / `invt`	Receivables (Channel stuffing risk) / Inventories
	`wcap`	Working Capital (Liquidity buffer)
	`ppent` / `intan`	Net PP&E / Intangibles (Soft assets)
	`gdwl`	Goodwill (Impairment risk)
	`txdba`	Deferred Tax Asset (Long Term) - *Proxy for NOLs*
Balance Sheet: Liab/Eq	`lt` / `lct`	Total Liabilities / Total Current Liabilities
	`dltt` / `dlc`	Long-Term Debt / Debt in Current Liabilities
	`dd1`	Long-Term Debt Due in 1 Year (*Refinancing wall*)
	`ap` / `txp`	Accounts Payable / Income Taxes Payable
	`txditc`	Deferred Taxes & Inv. Tax Credit (Non-current)
	`seq` / `re`	Stockholders' Equity / Retained Earnings (*Accum. Deficit*)
	`pstk` / `mib`	Preferred Stock / Noncontrolling Interest

Table 2: Partial table of accounting variables.

Expected Contribution

- **Replication:** Confirming the methodology and results of the working paper by Tewari et al. (2024) with the CRSP and Compustat Data.
- **Extension:** Extending the results of Tewari et al. (2024) by including autoencoders for extraction of the informational content in the features and by using the model-predictions for index construction.
- **Challenging traditional risk-scaling approaches:** Showing that ML applications are better suited to distinguish between Ecstasy and Agony stocks compared to classical risk-scaling methods, like Low-Volatility, Low-Beta or Altman's Z-score. Autoencoders and ensemble methods are well suited for capturing the complex interactions between accounting and macro variables.

Appendix: Frequency of Firms by CSI Count

Total CSI Events	Number of Firms	Percentage (%)
0	2167	66.41
1	537	16.46
2	255	7.81
3	135	4.14
4	93	2.85
5	51	1.56
6	18	0.55
7	7	0.21

Table 3: Number of observations per CSI count.

Appendix: Annual Frequency and Percentage of CSI Events

Year	Events	%	Year	Events	%
1998	83	3.71	2011	36	1.61
1999	56	2.50	2012	80	3.58
2000	124	5.55	2013	86	3.85
2001	16	0.72	2014	81	3.62
2002	57	2.55	2015	95	4.25
2003	44	1.97	2016	134	5.99
2004	22	0.98	2017	155	6.93
2005	60	2.68	2018	56	2.50
2006	69	3.09	2019	185	8.27
2007	68	3.04	2020	182	8.14
2008	55	2.46	2021	211	9.44
2009	114	5.10	2022	97	4.34
2010	70	3.13			

Table 4: Number of CSI events per year.

Bibliography

Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968.

Michael Cembalest. The agony & the ecstasy: The risks and rewards of a concentrated stock position. Eye on the market special edition, J.P. Morgan Asset Management, 2014. September 2, 2014.

Stephen Penman and Francesco Reggiani. Fundamentals of value versus growth investing and an explanation for the value trap. *Financial Analysts Journal*, 74 (4):103–119, 2018. doi: 10.2469/faj.v74.n4.6.

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