Vienna University of Economics and Business

Master Thesis Research Proposal

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

Author:
Tristan Leiter

Submission Date:
January 25, 2026

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

Chapter 1

Problem Description

1.1 Introduction

Research by J.P. Morgan Asset Management highlights an empirical phenomenon referred to as "The Agony and the Ecstasy" (Cembalest, 2014, 2024) revealing that equity indices, like the Russell 3000, are overwhelmingly influenced by extreme stock performances. While a small percentage of winners contribute the vast majority of excess returns, approximately 40% of all constituents suffer a "catastrophic decline," which is defined as a drawdown of 70% or more from their peak without a subsequent recovery.

Recent literature formalizes this phenomenon as a **"Catastrophic Stock Implosion"** (CSI) (Tewari et al., 2024). Unlike standard volatility, an implosion represents a distinct market event characterized by a severe price downturn followed by prolonged stagnation and minimal probability of recovery. This presents a critical challenge for index construction: passive investing captures the "Ecstasy" of winners but systematically forces investors to hold the "Agony" of these imploding assets (at least until removal from the constituent index). Crucially, as explored by Cembalest (2014, 2024), these declines are not limited to speculative "junk" companies; a significant portion of losers display "healthy" fundamentals prior to collapse, suggesting traditional metrics fail to capture the dynamics preceding permanent capital decline.

1.2 Status Quo

1. **The Quality Trap:** Recent literature has explored indicators which are perceived to be synonymous with healthy stock fundamentals. Penman and Reggiani (2018) suggest that the Book-to-Price ratio (B/P) is misleading. Low B/P values reflect uncertainty about future cash-flows rather than "cheap" buying opportunities. Additionally, Altman et al. (2016) argue that, amongst others, profitability is time-varying and find low predictive ability for longer time-periods. Specifically, while they find that profitability ratios like Return on Assets (ROA) provide efficient accuracy for a short horizon of two years , these measures fail to be consistent predictors over a ten-year horizon. They observe that in multivariate models, profitability is rendered largely insignificant when tested against solvency measures, such as the Equity Ratio, which dominates the prediction of distress irrespective of the horizon length.

2. **Predictive Ability:** Traditional bankruptcy models heavily relied on linear combinations of ratios, like the Z-score introduced by (Altman, 1968). Since then, a variety of risk-management models have been introduced, culminating in modern ML applications, such as Random Forest (RF) or support vector machines (Barboza et al., 2017). Even though logit/probit models work reasonably well, Jones et al. (2017) also recommends the use of more advanced machine-learning methods, like AdaBoost or RF.

3. **Misalignment of Prediction Horizons and Objectives:** While traditional bankruptcy models aim to minimize credit risk, they often fail to minimize market risk. A fundamental disconnect exists between *legal insolvency* and *market implosion.*

   - **Lagging Indicators:** Legal bankruptcy is frequently the final stage of a long deterioration process. By the time a traditional Altman Z-Score or structural model flags a company as distressed, the market has often already priced in the failure, resulting in a "Zombie state" where the asset lingers at depressed valuations (Tewari et al., 2024). For an equity investor, the capital is lost at the *implosion* event, not the bankruptcy filing.

- **The Cost of False Positives (Type I Errors):** In the context of equity indexing, the cost of a False Positive is mainly the opportunity cost. As noted in the "Agony and Ecstasy" framework, index returns are driven by a small tail of extreme winners. Traditional models, which penalize negative skewness too aggressively, risk flagging volatile but successful growth stocks as "distressed." Excluding a future "Ecstasy" stock (like NVIDIA) due to a conservative model, which produces too many False Positives, would severely underperform the benchmark, negating the benefits of avoiding the "Agony" stocks.

1.3 Research Gap and Proposed Methodology

1.3.1 Implications of Current Limitations

Ben Jabeur et al. (2021, 2023) have demonstrated that ML models, such as Gradient Boosting and Neural Networks, outperform statistical models in predicting financial distress by capturing non-linear relationships between variables. Tewari et al. (2024) have built on these insights and applied modern ML-techniques to market-based risk-modelling, for which they find that XGBoost can predict up to 61% of implosions in the test set with a false positive rate of less than 3%.

While the literature has identified a variety of ML-methods for credit-risk modelling, there is a lack of research on applying modern ML techniques to the specific problem of *market-based* catastrophic stock declines. Ensemble methods, like bagging and boosting, or Neural Networks are well suited to not only improve predictive accuracy, but also to capture the time-series dynamics of individual stock risk.

1.3.2 Proposed Approach: The "Crash-Filtered" Index

This thesis proposes bridging the gap between distress prediction and active index construction. While Tewari et al. (2024) established the concept of Catastrophic Stock Implosion (CSI), it remains unknown whether these insights can be operationalized into a viable risk-mitigation strategy. To address the "Agony and Ecstasy" dilemma, this research moves from "pure volatility forecasting" to "probabilistic implosion modeling" through the following three-stage methodology:

1. **Model Fitting and Selection:**
   The research will model the probability of market-risk loss, specifically the CSI event, by adopting the definition established by Tewari et al. (2024). The methodology extends current literature by moving beyond static snapshots of financial health. In addition to advanced ensemble methods (e.g., XGBoost, CatBoost) capable of handling non-linear interactions, this approach incorporates autoencoders to decode the noisy financial features.

2. **Index Construction:**
   Based on the model outputs, a "Crash-Filtered" equity index will be constructed. This process involves a systematic re-weighting or exclusion mechanism that penalizes index constituents exhibiting a likelihood of implosion exceeding a calibrated threshold. The primary objective is to penalize identified "Agony" candidates (permanent capital loss or capital loss exceeding a threshold) while retaining the "Ecstasy" winners (usually featuring high volatility and returns). This distinction is critical to isolating the alpha generated purely by tail-risk mitigation, rather than by a generic low-beta factor tilt.

3. **Backtesting and Performance Evaluation:**
   The efficacy of the strategy will be validated through out-of-sample backtesting using a rolling-forward cross-validation framework. The performance of the "Crash-Filtered" index will be benchmarked against:

   - The unfiltered market-weighted benchmark (e.g., Russell 3000) to test for alpha generation.
   - Traditional volatility-weighted portfolios (e.g., Minimum Volatility indices) to test for superior drawdown characteristics.

   Evaluation metrics will prioritize risk-adjusted returns and tail-risk characteristics.

Chapter 2

Research Question

Based on the identified problem that standard metrics fail to distinguish between "recoverable volatility" and "permanent implosion" and that perceived quality-signals can be misleading, the following research question could be explored:

2.1 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?

2.1.1 Sub-Question 1: Autoencoder

Standard ratios (like P/E) are too noisy, but an Autoencoder can find the "hidden" structure of a failing firm.

Does the integration of latent features derived from Denoising Autoencoders significantly improve the predictive performance (Average Precision) of Ensemble models compared to those trained solely on raw financial data?

Hypothesis: Autoencoders will successfully denoise volatile market indicators, allowing the Ensemble model to identify "structural" distress earlier than models relying on raw accounting inputs.

2.1.2 Sub-Question 2: The "False Positive" Advantage of Ensemble Methods

Showing that decision trees (handling the non-linearities) are better than the linear Altman Z-Score at saving the "Ecstasy" stocks.

Do Ensemble methods exhibit a superior ability to distinguish between recoverable volatility ('Ecstasy') and terminal implosion ('Agony') compared to indiscriminate volatility-based exclusion strategies?

Hypothesis: Ensemble models will distinguish between 'good' volatility (growth) and 'bad' volatility (implosion) more effectively than linear models, thereby reducing the exclusion of high-performing winners.

2.1.3 Sub-Question 3: Sensitivity

Does shortening the prediction horizon from 12 months to 6 months significantly enhance the index's ability to react to distress signals, or does it introduce excessive turnover without performance gains?

Prediction Horizon: Reacting faster to changes.

Chapter 3

Research Design

1. **Unsupervised Feature Extraction (Autoencoder)**
   To address the high dimensionality of the dataset (over 400 features, including TSFEL time-series aggregations and fundamental ratios like `ff_earn_yld` and `ff_cf_sales`), an Undercomplete Denoising Autoencoder is employed.

   The Autoencoder serves two distinct functions in the predictive pipeline:

   (a) **Dimensionality Reduction:** It compresses the noisy 400-dimensional input vector x into a lower-dimensional latent representation vector z (bottleneck layer). This extracts non-linear structural dependencies between variables (e.g., the interaction between decreasing Cash Flow from Operations and volatile Log Returns) that linear PCA would miss.

   (b) **Anomaly Detection Feature:** Following the future work suggested by Tewari et al. (2024), the model calculates the *Reconstruction Error* ($\mathcal{L} = ||x - \hat{x}||^2$). A high reconstruction error serves as a proxy for "abnormal" market behavior, added as an explicit feature to the supervised Ensemble model.

2. **Defining the target variable (Forward-Looking)**
   Consistent with the supervised learning framework of Tewari et al. (2024), the model is trained to predict the onset of a Catastrophic Stock Implosion (CSI).

   A stock is classified as a positive CSI case if it satisfies the following three conditions:

   (a) **Initial Crash:** The cumulative return drops below a threshold $C = -0.8$ (i.e., an 80% drawdown from the trailing peak).

   (b) **Zombie Period:** The stock enters a stagnation period of duration $T = 78$ weeks (approx. 1.5 years).

   (c) **Non-Recovery:** Throughout the zombie period, the cumulative return never exceeds the recovery ceiling $M = -0.2$ (remaining at least 20% below the peak).

   While Tewari et al. (2024) employ a yearly prediction horizon ($h = 12$ months), this study follows this approach and also tests whether this horizon can be relaxed to $h = 6$ months to enhance the temporal precision required for more frequent index rebalancing. The binary target variable $y_{i,t}$ is defined as:

$$y_{i,t} = \begin{cases} 1 & \text{if stock } i \text{ triggers } C = -0.8 \text{ within } [t, t+h] \text{ and satisfies the zombie criteria} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

   Sensitivity Analysis and Parameter Relaxation:
   The baseline definition of a 1.5-year ($T = 78$ weeks) zombie period is highly conservative. To assess robustness, this study will perform a sensitivity analysis by relaxing the duration parameter T to **less than 1 year** (e.g., $T = 26$ or $T = 52$ weeks). This tests whether the "Agony" of implosion can be effectively captured with a shorter confirmation window, potentially allowing the model to identify distress signals earlier without significantly increasing the False Positive Rate.

3. **Model Development**

   - **Algorithm:** Ensemble methods (XGBoost, CatBoost) are used to predict the probability of CSI ($y_{i,t} = 1$).

- **Validation of Subquestion 1:** To isolate the contribution of the Autoencoder, an internal comparison will be conducted between:
  - *Model A:* Gradient Boosting on Raw Financial Ratios.
  - *Model B:* Gradient Boosting on Autoencoder Latent Features.
- **Validation of Subquestion 2:** The best-performing Ensemble model is compared against:
  - The Naive Baseline (class prior).
  - Logistic Regression.
  - The **Inverted Altman Z-Score**, treated as a continuous risk ranking to allow for direct AUC-PR comparison.

4. **Index Construction**
   The "Crash-Filtered" index is constructed by filtering the CRISP-universe.

   - **Rebalancing:** The portfolio is rebalanced **anually** or **semi-anually** (dependent on the prediction horizon, e.g. $h = 12$ months) to ensure timely removal of deteriorating assets.
   - **Filtering Mechanism:** For each stock, if the predicted probability $\hat{p}_{i,t} > \theta$, it is excluded.
   - **Calibration:** θ is calibrated on the validation set to maximize the F1-score (harmonic mean of Precision and Recall), balancing the cost of False Negatives (holding an implosion) against False Positives (missing a winner).

5. **Performance Evaluation**
   The strategy is backtested using a rolling-forward cross-validation framework. Performance is benchmarked against the **Market-Weighted Index** (unfiltered) and the **MSCI-USA Minimum Volatility Index**.

Chapter 4

Expected Contribution

1. **Methodological Advancement: A Hybrid Unsupervised-Supervised Framework**
   While Tewari et al. (2024) successfully demonstrated the efficacy of supervised learning (XGBoost) for predicting catastrophic implosions, they explicitly identified unsupervised learning as a "promising avenue for research" to model implosions as anomalies. This thesis directly addresses this call by introducing a **Hybrid Autoencoder-Ensemble architecture**. By using an Autoencoder to denoise high-dimensional financial data and extract latent "distress structures," this research contributes a novel feature engineering pipeline that aims to improve predictive performance in noisy market regimes where standard ratios fail.

2. **Practical Operationalization: From Prediction to Investment Utility**
   Existing literature often focuses on the statistical accuracy of distress models (e.g., AUC, Precision, Recall). However, a statistical prediction does not guarantee economic value. This thesis shifts the focus from "pure prediction" to "portfolio construction." By developing a systematic **"Crash-Filtered" Index**, this research validates whether the statistical signal of a Catastrophic Stock Implosion (CSI) can be monetized. It provides empirical evidence on whether avoiding specific "Agony" events generates superior risk-adjusted returns compared to generic Minimum Volatility strategies, thus bridging the gap between machine learning metrics and investor outcomes.

3. **Theoretical Distinction: Disentangling Volatility from Implosion**
   Traditional risk models treat volatility as a proxy for risk, often penalizing high-growth "Ecstasy" stocks that exhibit "good" volatility. This thesis contributes to the "Agony and Ecstasy" framework (Cembalest, 2014) by empirically testing whether "Catastrophic Implosion" is a distinct risk factor separate from standard price variance. By calibrating the model to minimize Type I errors (False Positives), this research seeks to demonstrate that it is possible to decouple downside tail-risk protection from the upside participation of growth stocks—a distinction that standard low-beta factors fail to achieve.

Chapter 5

Refinements

Table 5.1: Long-Term Growth Distribution: Imploded vs. Non-Imploded Firms

	Cohort Distribution	
Growth_Category	Imploded Firms	Never Imploded
High Growth (>10%)	4.9%	40.5%
Moderate Growth (5-10%)	9.1%	25.9%
Low Growth (0-5%)	13.8%	13.7%
Low Losses (0 to -5%)	13.0%	7.0%
High Losses (<-5%)	59.1%	12.9%

Table 5.2: Long-Term Growth Distribution by Frequency of Implosion Events

	Number of CSI Events Experienced							
Growth_Category	Events: 0	Events: 1	Events: 2	Events: 3	Events: 4	Events: 5	Events: 6	Events: 7
High Growth (>10%)	40.5%	7.4%	3.1%	3.0%	1.1%	0.0%	5.6%	0.0%
Moderate Growth (5-10%)	25.9%	10.8%	8.6%	7.4%	9.7%	2.0%	0.0%	0.0%
Low Growth (0-5%)	13.7%	13.6%	16.5%	11.9%	14.0%	9.8%	11.1%	0.0%
Low Losses (0 to -5%)	7.0%	10.1%	12.5%	19.3%	19.4%	15.7%	22.2%	14.3%
High Losses (<-5%)	12.9%	58.1%	59.2%	58.5%	55.9%	72.5%	61.1%	85.7%

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 (<-2%)	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.

1. **Validation of the "Avoidance Alpha" Hypothesis**

   The analysis confirms that the CSI signal successfully isolates "toxic" assets. In the "Imploded" cohort, **48.5%** of firms resulted in long-term *Value Destruction* (CAGR $< -2\%$), compared to only **16.8%** in the "Clean" cohort.

   This 3x increase in the probability of value destruction validates the conservative definition of the zombie period (T = 78 weeks). It demonstrates that the target variable $y = 1$ is not capturing noise, but rather

Table 5.3: Long-Term Growth Distribution by Frequency of Implosion Events

	Number of CSI Events Experienced							
Growth_Category	Events: 0	Events: 1	Events: 2	Events: 3	Events: 4	Events: 5	Events: 6	Events: 7
High Growth (>10%)	40.5%	7.4%	3.1%	3.0%	1.1%	0.0%	5.6%	0.0%
Moderate Growth (5-10%)	25.9%	10.8%	8.6%	7.4%	9.7%	2.0%	0.0%	0.0%
Low Growth (2-5%)	9.0%	8.4%	8.6%	8.1%	8.6%	3.9%	11.1%	0.0%
Stagnation (-2-2%)	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Value Destruction (<-2%)	15.4%	60.3%	65.9%	68.9%	66.7%	86.3%	72.2%	85.7%
Value Destruction (No Recovery)	0.8%	3.5%	2.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Unknown	8.4%	9.5%	11.8%	12.6%	14.0%	7.8%	11.1%	14.3%

a fundamental deterioration in solvency. Consequently, the primary mechanism of the "Crash-Filtered" Index is empirically justified: the outperformance is expected to stem primarily from the *avoidance of zeros* (insolvency) rather than the selection of winners.

2. **Quantifying the "Phoenix Cost" (Type I Error Trade-off)**
   A critical refinement for the rebalancing strategy arises from the finding that **12.5%** of Imploded firms eventually became "High Growth" assets ($>$ 10% CAGR). These are "Phoenix" stocks that suffered a catastrophic drawdown but successfully restructured (e.g., Amazon in 2001).

   This finding highlights a distinct "Phoenix Cost"—the opportunity cost of excluding a turnaround winner. To mitigate this, the Research Design (Chapter 3) must explicitly incorporate a **Re-entry Mechanism**. The index cannot permanently blacklist a firm following a CSI event. Instead, the annual or semi-annual rebalancing frequency (h = 6 or 12 months) is essential to allow the model to reassess the risk. Once the Autoencoder's anomaly signal decays, the strategy must permit these "Phoenix" stocks to re-enter the portfolio to capture the post-distress recovery phase.

3. **The "Serial Offender" Non-Linearity**
   The breakdown of outcomes by event count reveals a non-linear dose-response relationship. Firms experiencing exactly one CSI event show a **47.7%** rate of value destruction, which rises to **53.1%** for firms with two events.

   This refines the feature engineering approach for the Supervised Model. It suggests that "Past CSI Count" or "Time Since Last Implosion" are not merely historical metadata but predictive state variables. A linear model might interpret a -80% crash as a "mean reversion" buying opportunity; however, the data suggests that conditional on crashing once, the probability of future stagnation remains structurally elevated. This justifies the use of the **Autoencoder**, which is capable of encoding this non-linear "memory" of distress into the latent feature vector z, differentiating between a "cheap" value stock and a "structurally impaired" serial offender.

Chapter 6

Data

6.1 Constituent Universe (CRSP)

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.

6.1.1 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 **??**).

Table 6.1: Aggregate Sample Statistics

Statistic	Value
Total Unique Firms (N)	3,263
Total Observations	51,773
Average Years per Firm	14
Sample Start Date	1998-12-31
Sample End Date	2024-12-31
Total Identified Implosions	2,236

6.1.2 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 6.2: 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

6.2 Features (Compustat)

To capture corporate fundamental signals, we utilize the **Compustat Fundamentals Annual** database (`comp.funda`) accessed via WRDS. This dataset provides comprehensive balance sheet, income statement, and cash flow data for North American publicly traded companies.

6.2.1 Data Linking and Merging

To align fundamental data with the market data defined in the previous section, we employ the **CRSP/Compustat Merged (CCM)** linking table (`crsp.ccmxpf_lnkhist`). This ensures accurate mapping between CRSP permanent identifiers (`permno`) and Compustat identifiers (`gvkey`). The linking process enforces the following strict validity criteria:

- **Link Type:** Restricted to `LC` (Link Research), `LU` (Unresearched Link), and `LS` (Link System) to ensure high-confidence matches.
- **Link Primary Marker:** Limited to `P` (Primary) and `C` (Composite) to isolate the primary trading vehicle for the firm.
- **Date Validity:** Fundamental observations are only retained if the fiscal year-end date falls within the valid range defined by the link start (`linkdt`) and end dates (`linkenddt`).

6.2.2 Filtering and Universe Alignment

Within the Compustat universe, we apply standard filters to ensure data consistency and comparability across firms. We restrict the sample to the Industrial Format (`indfmt="INDL"`) and Standardized Data Format (`datafmt="STD"`). Furthermore, we utilize Consolidated financial statements (`consol="C"`) to capture the full economic entity of the firm.

6.2.3 Look-Ahead Bias Mitigation

TBD.

6.2.4 Feature Selection

We extract a comprehensive set of fundamental variables capturing distinct dimensions of firm health. Beyond standard performance metrics, our selection specifically targets **distress precursors** identified in the "Zombie" firm literature. These include *Retained Earnings* (to capture historical accumulated deficits), *Deferred Tax Assets* (proxying for loss carryforwards), and *Discretionary Expenses* (advertising and R&D cuts signaling cash preservation).

The feature set is organized into three primary categories: Balance Sheet Structure, Operational Performance, and Supplementary Distress Signals. The complete list of variables is detailed in Table 6.3.

Table 6.3: Selected Fundamental Features (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
	`tstk`	Treasury Stock (Contra-equity)
Income Statement	`sale` / `gp`	Net Sales / Gross Profit
	`cogs` / `xsga`	Cost of Goods Sold / SG&A Expense
	`ebit` / `oibdp`	EBIT / Operating Income Before Deprec. (*EBITDA*)
	`ni` / `pi`	Net Income / Pretax Income (*Tax efficiency check*)
	`epsfi`	Earnings Per Share (Basic)
	`spi`	Special Items (*Early warning signal*)
	`xint` / `xinst`	Total Interest Exp. / Short-Term Interest Exp.
	`xrd` / `xad`	R&D Exp. / Advertising Exp. (*Discretionary cuts*)
	`xrent`	Rental Expense (Operational rigidity)
	`dp`	Depreciation & Amortization
Supplementary	`emp`	Number of Employees (*Real economy anchor*)
	`dpr` / `ppegt`	Accum. Depreciation / Gross PPE (*Asset age ratio*)
	`fca`	Foreign Exchange Income/Loss
	`mkvalt`	Market Value (Fiscal Year-End)

6.3 Dataset Merging

Chapter 7

Descriptive Statistics

7.1 Dataset

7.1.1 Sample Statistic

Table 7.1: Aggregate Sample Statistics

Statistic	Value
Total Unique Firms (N)	3,263
Total Observations	51,773
Average Years per Firm	14
Sample Start Date	1998-12-31
Sample End Date	2024-12-31
Total Identified Implosions	2,236

Table 7.2: Distribution of Catastrophic Stock Implosion (CSI) Events per Firm

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

7.1.2 Industry Distribution

Table 7.3: Temporal Distribution of Catastrophic Stock Implosions (Target Variable y)

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

Table 7.4: Constituent Distribution by Industry Sector (NAICS/SIC Hybrid)

Industry Sector	Count	Percentage (%)	Source
Manufacturing	1169	35.9	NAICS
Finance and Insurance	510	15.7	NAICS
Real Estate and Rental and Leasing	214	6.6	NAICS
Professional, Scientific, & Technical Services	191	5.9	NAICS
Information	188	5.8	NAICS
Mining, Quarrying, and Oil & Gas Extraction	156	4.8	NAICS
Admin. Support & Waste Mgmt	127	3.9	NAICS
Retail Trade	126	3.9	NAICS
Wholesale Trade	113	3.5	NAICS
Transportation and Warehousing	96	2.9	NAICS
Utilities	72	2.2	NAICS
Management of Companies and Enterprises	69	2.1	NAICS
Accommodation and Food Services	62	1.9	NAICS
Health Care and Social Assistance	56	1.7	NAICS
Construction	53	1.6	NAICS
Arts, Entertainment, and Recreation	14	0.4	NAICS
Other Services (except Public Admin)	13	0.4	NAICS
Agriculture, Forestry, Fishing & Hunting	12	0.4	NAICS
Educational Services	12	0.4	NAICS
Public Administration	1	0.0	NAICS
NA	1	0.0	SIC (Proxy)

Chapter 8

Objective Function

8.1 Model Selection and Optimization Strategy

8.1.1 The Economic Case for Recall at Fixed False Positive Rates

The selection of an appropriate objective function is critical when bridging the gap between statistical probability and investment utility. While standard classification metrics like Accuracy or F1-Score are ubiquitous, they often fail to capture the asymmetric costs inherent in the "Agony and Ecstasy" framework. In the context of constructing a "Crash-Filtered" Index, the cost of a False Positive (Type I Error) is not merely a statistical nuisance but a substantial opportunity cost: identifying a high-growth "Ecstasy" stock (e.g., NVIDIA) as a crash candidate results in its exclusion, thereby generating underperformance relative to the benchmark.

Consequently, this thesis adopts a two-staged evaluation strategy:

1. **Hyperparameter Optimization (CV): Average Precision (AP).** During the cross-validation and training phase, we optimize for the Area Under the Precision-Recall Curve (AUC-PR/AP). Unlike the Receiver Operating Characteristic (ROC), AP focuses exclusively on the performance of the minority class (Catastrophic Stock Implosion). Since the dataset is highly imbalanced ($\approx 5\%$ prevalence), AP provides a more robust signal for tuning hyperparameters (e.g., tree depth, learning rate) without committing to a specific decision threshold, ensuring the model effectively ranks distress probabilities.

2. **Final Model Selection: Recall at Fixed FPR.** For the final economic evaluation, we transition to a constraint-based metric: **Recall at Fixed False Positive Rate (FPR)**. This metric aligns directly with the practical "Risk Budget" of a portfolio manager. By fixing the FPR at conservative levels (e.g., 3%, 5%, or 10%), we effectively set a "budget" for mistakenly excluding healthy firms. The objective is to maximize the number of "Agony" stocks identified (Recall) while strictly capping the exclusion of potential "Ecstasy" winners (FPR). This approach empirically tests whether the model can decouple tail-risk protection from upside participation, a key limitation of traditional Minimum Volatility strategies.

8.2 The Potential Pitfall of AUC-ROC

8.2.1 The "False Positive" Trap in Imbalanced Datasets

While the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is the standard metric in bankruptcy prediction literature, it presents a potential "trap" for this specific research design. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate across all thresholds. In a dataset where negatives (non-imploding firms) comprise over 95% of observations, the FPR denominator is large. Consequently, a model can misclassify a significant number of high-volatility "Ecstasy" stocks as implosions (False Positives) without significantly increasing the calculated FPR, thereby inflating the AUC-ROC score.

This phenomenon is particularly dangerous for "Avoidance Alpha" strategies. A high AUC-ROC might reflect a model that is excellent at classifying obvious, low-volatility "safe" stocks (the easy negatives) but fails to distinguish between high-volatility winners and high-volatility losers in the tail.

Therefore, while this study will report AUC-ROC for comparability with existing literature (e.g., Altman (1968)), it serves as a secondary diagnostic. We explicitly monitor for divergence between AUC-ROC and Average Precision to ensure the model is not achieving high scores simply by overfitting to the majority class of "safe" firms while failing to solve the central "Agony and Ecstasy" dilemma.

Chapter 9

Autoencoders

9.1 Feature extraction

Autoencoders (AE) serve as powerful unsupervised learning tools for modeling complex financial data. In the context of predicting Catastrophic Stock Implosion (CSI), we identify two primary architectural strategies: Denoising Autoencoders for robust feature cleaning and Reconstruction Error for anomaly detection.

9.1.1 Architectural Strategies

Use Case A: Denoising Autoencoders (DAE) for Manifold Learning

The primary objective of a Denoising Autoencoder (DAE) in financial forensics is to recover the "true" economic health of a firm from "noisy" accounting data. Financial statements often contain noise due to earnings management, temporary accruals, or reporting inconsistencies.

- **Mechanism:** The DAE is trained to map a corrupted input vector, $\tilde{x}$, back to the original uncorrupted input, x. The corruption is introduced stochastically (e.g., via Gaussian noise or masking).
- **Objective Function:** The model minimizes the loss function $L(x, g(f(\tilde{x})))$, where f is the encoder and g is the decoder.
- **Theoretical Benefit:** By forcing the network to remove noise, the AE learns the underlying *manifold* of valid financial structures (e.g., the intrinsic relationship between Assets, Liabilities, and Equity).
- **Application:** The learned bottleneck representation serves as a robust, noise-reduced feature set for the downstream supervised classifier (e.g., predicting $y = 1$).

Use Case C: Anomaly Detection via Reconstruction Error

Catastrophic stock implosions are, by definition, anomalies. A standard Autoencoder can be utilized as a semi-supervised anomaly detector rather than just a feature extractor.

- **Mechanism:** An AE is trained exclusively on "healthy" firms (where $y = 0$) to learn the latent structure of non-imploding companies.
- **Inference:** When a potentially catastrophic firm (a future CSI event) is passed through this trained network, the AE will struggle to reconstruct the input because the firm's financial structure deviates from the healthy manifold.
- **New Feature:** The *Reconstruction Error* (typically Mean Squared Error between the input and the output) is calculated for each observation.
- **Application:** This error value is treated as a high-signal feature in the final supervised model. A high reconstruction error acts as a proxy for "structural financial abnormality."

9.1.2 Implementation in R

The following R packages and functions are recommended for implementing the strategies outlined above.

For Use Case A (Denoising)

The `keras` package is preferred for DAEs due to its flexibility in adding custom noise layers, which are essential for the denoising architecture.

- **Package:** `keras` (Interface to TensorFlow)
- **Core Functions:**
  - `layer_gaussian_noise()`: injects stochastic noise into the input layer during training.
  - `layer_dropout()`: alternative method to corrupt inputs by randomly zeroing out features.
  - `fit()`: utilized with distinct x (target) and $\tilde{x}$ (input) data to train the reconstruction task.

For Use Case C (Anomaly Detection)

The `h2o` package is highly efficient for this task, particularly given its ability to handle tabular accounting data and missing values natively without extensive preprocessing pipelines.

- **Package:** `h2o`
- **Core Functions:**
  - `h2o.deeplearning(..., autoencoder = TRUE)`: trains the AE model.
  - `h2o.anomaly()`: a specialized function that automatically calculates the reconstruction error (MSE) for every row in a new dataset based on the trained model.

Bibliography

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

Edward I. Altman, Małgorzata Iwanicz-Drozdowska, Erkki K. Laitinen, and Arto Suvas. Financial and non-financial variables as long-horizon predictors of bankruptcy. *Journal of Credit Risk*, 12(4):49–78, 2016.

Flavio Barboza, Herbert Kimura, and Edward I. Altman. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417, 2017.

Sami Ben Jabeur, Cheima Gharib, Salma Mefteh-Wali, and Wissal Ben Arfid. Catboost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166:120658, 2021.

Sami Ben Jabeur, Nicolae Stef, and Pedro Carmona. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Computational Economics*, 61:715–741, 2023.

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.

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, 2024.

Stewart Jones, David Johnstone, and Roy Wilson. Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting*, 44(1–2):3–34, 2017.

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.